IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 17, 2020, accepted January 23, 2021, date of publication January 28, 2021, date of current version February 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055433

A Gaussian Mixture Model Clustering Ensemble Regressor for Semiconductor Manufacturing Final Test Yield Prediction

DAN JIANG[1,2], **WEIHUA LIN**[2], **AND NAGARAJAN RAGHAVAN**[1], **(Member, IEEE)**

[1]Engineering Product Development, Singapore University of Technology and Design, Singapore 487372

[2]Silicon Laboratories International Pte. Ltd., Singapore 539775

Corresponding author: Dan Jiang (dan_jiang@mymail.sutd.edu.sg)

This work was supported by the Economic Development Board of Singapore (EDB) and Silicon Laboratories, Inc. (SLAB) through the Industry Postgraduate Program (IPP) under Grant IGIPSILICON1901.

ABSTRACT In the semiconductor industry, many studies have been carried out for front-end related process improvement and yield prediction using machine learning techniques. However, very few research investigations have dealt with the backend Final Test (FT) yield prediction using the front-end wafer acceptance test (WAT) parameters. The manufacturing cycle time between wafer fabrication (WF) and FT can range anywhere between a few weeks to several months. It is therefore important for semiconductor manufacturers to detect wafer material related low yield problems at an earlier stage for effective cost and quality control. This is a challenging goal as the input data used for prediction is at a very early manufacturing stage and the output FT yield for packaged chips is the last stage of the fabrication chain. There are many unknown production variations caused by different manufacturing processes, equipment configurations and human interferences in this multi-stage sequential fabrication chain. In this paper, we proposed a novel procedure to predict the backend FT yield at the WF stage itself using a Gaussian Mixture Models (GMM) clustering approach that is applied to build a weighted ensemble regressor. Real production data for new chip product lines are verified with this method and show significant improvement in the prediction performance.

INDEX TERMS Semiconductor manufacturing, yield prediction, final test, Gaussian mixture models, clustering, regression, ensemble methods, smart manufacturing.

I. INTRODUCTION

In today's competitive semiconductor industry where huge amount of data are generated every day from hundreds to thousands of manufacturing process steps, advanced data analytics solutions are gaining increasing importance to improve capacity, quality and efficiency. In particular, production yield analysis is one of the most important areas of focus from a semiconductor device manufacturer's (foundry's) operational cost perspective. The typical semiconductor manufacturing process begins from the front-end all the way to the back-end and packaging where four major tests are conducted. The front-end process includes wafer fabrication (WF) and wafer probing. During WF, various test structures are fabricated on a wafer to extract information on the process and device performance for yield management [1]. Once WF is completed, the wafer acceptance test (WAT) is conducted using these test structures to measure important process related parameters, such as contact resistance, threshold voltage and diode leakage current etc. The size of the test structures used depends on the semiconductor technology node. In general, only less than 10 test structures for measurements are sampled from each wafer. It makes WAT data very difficult to analyse because it only covers around 10% of the total dies on a single wafer. However, it is important to analyse WAT data, especially for fabless semiconductor companies because this is the first data available during the entire manufacturing flow which can be used for wafer quality evaluation and production forecast. The second test is wafer probing, which provides functional test coverage on all the dies. Following this is the back-end process which includes assembly and Final Test (FT). After assembly, each back-end lot goes through an assembly test to screen out continuity rejects. The FT involves a chip-level testing and it

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaojun Li.

has the largest test coverage in terms of device functionality. Therefore, the FT yield varies a lot depending on the process variation, test methodology and equipment condition. The FT yield is one of the major factors directly influencing the manufacturing operational cost. Low yield problem at the FT stage can drastically affect the company's gross margins. Current practice for low yield problem analysis is to first monitor the production FT yield. Once there is a low yield issue triggered at the back-end, the engineers trace back to the wafer front-end process manually and speculate the root cause. It can take several weeks or months to catch the front-end problem due to long production duration from WF to FT. Besides, design of experiment (DOE) is a common tool for wafer process related root cause investigation. However, DOE is only applicable for univariate and bivariate parameter analysis. Wafer manufacturing process consists of hundreds to thousands of parameters. Therefore, it is important to be able to develop a predictive tool which is capable of analysing high dimensional parameters to identify process related problems at a much earlier stage and facilitate corrective actions more effectively, resourcefully, and efficiently.

In this paper, we propose a novel FT yield prediction model using the Gaussian Mixture Model (GMM) clustering approach. The front-end WAT measurements are used as input data to predict the back-end FT test yield. This makes the problem more challenging compared to past research studies since the manufacturing process variation is not limited to the front-end process alone. The back-end process including assembly and FT methodology introduces more uncertainty into the predictive model.

The main motivation for our study is to have an automated yield prediction tool to identify low yield problems at a much earlier production stage compared to current practice. Besides, our tool is able to automatically identify and rank important WAT parameters and provide quick fix yield improvement strategies, thereby reducing (if not eliminating) the cost and time duration for DOE to be performed. In general, this approach will enable significant manufacturing cost reduction and help monitor product quality, as well as improve shipment forecast accuracy.

The remainder of the paper is organized as follows. Section II discussed related work in semiconductor domain. Section III introduces the machine learning methodologies used in this work. Section IV describes the flow of FT yield prediction procedure. Section V presents and discusses the results by using real production data from a couple of new chip product lines applied to the procedure. Section VI presents the case study for feature importance analysis and yield improvement validation. Finally, we conclude our study in Section VII along with possible suggestions for further work.

II. RELATED WORK

Previous semiconductor industry research explorations used machine learning techniques that were mainly focused on the front-end process related problems like virtual metrology (VM) improvements in Ref. [2], fault detection and classification in Ref. [3], wafer yield estimation in Refs. [4], [5], probe yield excursion detection and root cause analysis in Ref. [6], probe yield analysis based on wafer spatial features in Refs. [7], [8] and probe yield prediction with input parameters including electrical test parameters, wafer defect and wafer physical data [9], [10]. Based on our extensive survey, there are limited studies for back-end yield related problems. One such study by Park *et al.* in Ref. [11] demonstrates a framework to predict the FT yield based on probe test parametric data. Another work by Kang *et al.* in Ref. [12] talks about a FT yield classifier which is proposed by using wafer probe test results and wafer map features as the input.

To the best of our knowledge, there is no study that deals with direct backend FT yield prediction using the WAT parameters at the initial WF stage, which is the key motivation of our study here. In general, there are two major common difficulties for semiconductor manufacturing yield prediction problems, which are high dimensional input data and complex process variations. To tackle the high dimensional input data problem, the Pearson correlation is one of the common feature reduction methods mentioned in Ref. [5]. Mutual information (MI) and recursive feature elimination (RFE) are also popular feature selection techniques applied in Ref. [13] for manufacturing cycle time (CT) prediction and in Ref. [14] for etching process fault detection. A novel feature selection method for identifying the key parameters of WAT measurements based on Hybrid Feature Selection (HFS) was proposed by Xu *et al.* in their work of Ref. [15]. The HFS method can effectively filter out the noise parameters and achieve accurate prediction of wafer probe yield with reduced key WAT parameters. However, this method is computationally expensive because the genetic algorithm used requires a considerably large number of iterations to provide a stable solution. Besides, the method is not able to provide the WAT parameters' importance ranking using the generated deep belief network (DBN) model. Deep Neural Network (DNN) is a popular method for wafer map related studies in Refs. [7], [8]. However, DNN is not able to directly provide feature importance analysis. The over-fitting problem and poor model visibility as mentioned in Ref. [9] makes DNN not suitable in our study where being able to do low yield root cause analysis is one of the key priorities.

In Ref. [16], a regression-based model was proposed for CT factor selection by continuous factors discretization and stepwise CT-related factor selection. This model is more suitable for applications where the input parameters are a mixture of numerical and categorical data. Therefore, it is unfit for our case when input data are only WAT parameters and all the more, discretization will reduce the input information. Wang *et al.* in Ref. [17] introduced a factor selection algorithm for CT explanatory network combined with MI and network deconvolution techniques. Their results indicate that the proposed method has higher effectiveness to identify the explanatory variables. This algorithm is more suitable for neural network-based algorithms with data sets containing both direct and indirect dependencies.

The other major difficulty is the complex phase-wise manufacturing process which renders the production data unsuitable for most of the machine learning model assumptions. Ideally, one would prefer that the input and output parameter distributions follow the Gaussian distribution; but in real production scenarios, the parameter distributions exhibit clustering effects and high skewness due to process variations, human interference and equipment performance instability or degradation. The root cause for the low yield issue for each of the identified clusters can be quite different. A study by Pampuri *et al.* in Ref. [2] developed a multilevel lasso model, showing that the L1 penalized machine learning technique is suitable to handle data heterogeneity caused by inhomogeneous production and equipment logistics. However, this model is not generalizable because it requires a sound understanding of the detailed process to design such a model.

In Ref. [18], Chen discussed using the principal component analysis (PCA) approach to enhance the forecasting performance of the fuzzy back propagation network (FBPN) for WF CT estimation. In his study, PCA was applied to formulate variables that are independent of each other and thereby become new inputs to the FBPN. This approach is not suitable here for FT yield prediction because root cause analysis is important for yield improvement and using the PCA for feature selection will cause WAT parameters' information loss and the independent variables become less interpretable and controllable. A bi-directional classifying fuzzy-neural approach was also introduced by Chen in Refs. [19], [20] for WF CT estimation. The author applied Fuzzy c-mean (FCM) clustering for job CT pre-classification using time related parameters as the input. However, this approach also does not apply to FT yield clustering because the yield distribution tends to be mixture of Gaussians and FCM performs well only in the case of clustering of spherical clusters as discussed by Suganya *et al.* in Ref. [21]. Besides, FCM is sensitive to noisy data [21] and experiences difficulty in handling outlier points as mentioned by Thomas *et al.* in Ref. [22].

Grid search and manual search are the most widely used strategies for hyper-parameter optimization of machine learning algorithms as discussed in Ref. [23]. Grid search was applied to improve Support Vector Regressor (SVR) prediction performance for VM in Ref. [24]. In our work, the grid search strategy is used for model optimization due to limited data size and also because there are no neural network based models under consideration here for model selection.

Ensemble method is used to improve the prediction accuracy through combining several models. Saqlain *et al.* proposed a voting ensemble classifier with multi-type features to identify wafer map defect patterns in Ref. [25]. The ensemble classifier was also applied in Ref. [26] to improve accuracy of wafer failure map pattern classification. The main idea of clustering is to group the models in several clusters and choose representative models (one or more) from each cluster [27]. In this study, we use the cluster ensemble method, which combines multiple partitionings of a set of objects without accessing the original features (refer to Ref. [28]). The purpose of it is not limited to only improve prediction performance, but more importantly, to exploit the important and implicit WAT parameters within the sub-clusters for yield improvement purpose.

III. METHODOLOGIES TO BE CONSIDERED

A. FEATURE SELECTION METHODS

1) MUTUAL INFORMATION FEATURE SELECTION

MI is a measure of variable similarity between random variables, and it describes the difference between the entropy and conditional entropy for two random variables X and Y. The definition is

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \tag{1}$$

where $P(x)$ and $P(y)$ represent the marginal distributions of the data points in X and Y, and $P(x, y)$ is the joint distribution. A lower value of the MI feature implies that the variables are more independent.

2) RECURSIVE FEATURE ELIMINATION

Recursive Feature Elimination (RFE) refers to model building using the entire dataset and presentation of ranking of the feature importance. The least significant feature at each iteration is removed. The whole process is repeated until no more improvements are observed in the prediction performance. For simplicity, in this paper, we use a basic ordinary least square (OLS) algorithm to fit the dataset. The *F-statistic* and *p-values* are used for feature ranking.

B. GAUSSION MIXTURE MODELS (GMM)

A Gaussian Mixture Model (GMM) is a probabilistic model representing a mixture of a finite number of Gaussian distributions with unknown weights, means and covariances. The Expectation-Maximization (EM) algorithm is used to estimate the parameters in the GMM. Assume we have a K component Gaussian mixture, given training data set, $X = (x^1, \ldots x^N)$ and joint distribution of x with a latent variable z given by:

$$p(x,z) = p(x|z)p(z) \tag{2}$$

Let π_k be the mixture weights for the k components and therefore

$$p(z=k) = \pi_k, \quad k = 1, \ldots K \tag{3}$$

The overall joint probability of X with the latent variable Z is then given by:

$$P(X,Z) = \sum_{k=1}^{K} \pi_k N(X_i|\mu_k, \Sigma_k) \tag{4}$$

In order to get the values of the parameters π_k, μ_k, Σ_k, which represent the mixture weights, means and covariances between the Gaussian distributions, we solve for the maximum log-likelihood function iteratively applying the following two steps in the EM algorithm till convergence is achieved.

1) EXPECTATION STEP

The likelihood is defined by:

$$L(\pi, \mu, \Sigma) = \sum_{i=1}^{N} \log(z_k^i) \sum_{k=1}^{K} p(x^i|z=i)p(z=k) \qquad (5)$$

By resolving Eqn. (5), we have the posterior probability for component z_k^i at iteration $t+1$ represented as:

$$\gamma_{t+1}(z_k^i) = P_{\pi(t),\mu(t),\Sigma(t)}(z=k|x^i) \qquad (6)$$

2) MAXIMIZATION STEP

We update our estimate of the mixture weight, mean and covariance of each Gaussian cluster at the iteration, $t+1$, with:

$$\pi_k(t+1) = \frac{1}{N}\sum_{i=1}^{N} \gamma_{t+1}(z_k^i) \qquad (7)$$

$$\mu_k(t+1) = \frac{\sum_{i=1}^{N} x^i \gamma_{t+1}(z_k^i)}{\sum_{i=1}^{N} \gamma_{t+1}(z_k^i)} \qquad (8)$$

$$\Sigma_k(t+1) = \frac{\sum_{i=1}^{N} x^i \gamma_{t+1}(z_k^i)(x^i - \mu_k(t+1))(x^i - \mu_k(t+1))^T}{\sum_{i=1}^{N} \gamma_{t+1}(z_k^i)} \qquad (9)$$

C. BAYESIAN INFORMATION CRITERION

The Bayesian information criterion (BIC) is used as an estimate of the Bayes factor for two or more competing models. It is a suitable criterion to choose the optimal number of clusters for GMM [29]. It is defined simply by:

$$BIC = \ln(n)d - \ln(\hat{L}) \qquad (10)$$

where n is the number of data points, d is the number of parameters in the model, and $\hat{L}$ is the maximized value of the fitted model likelihood. When L increases, the BIC score will decrease which implies a better fitted model. The first term of the BIC expression represents the penalty incurred due to over-fitting when too many parameters are in the model.

IV. YIELD PREDICTION PROCEDURE

The overall process for FT yield prediction includes four major steps: (A) Data pre-processing, (B) GMM clustering, (C) Feature selection and model selection, and (D) Model optimization and model ensemble. The flow chart of our entire yield prediction framework is presented in Fig 1.

A. DATA PRE-PROCESSING

For the first step of data pre-processing, we calculate the mean and standard deviation using past three years' production WAT parameters. The mean and standard deviation for each individual parameter is used for input data standardization (normalization). The purpose is to reduce yield prediction bias caused by different WAT parameters' scale of values. Besides, the Pearson correlation is used to remove highly correlated parameters with a $p-value$ larger than 0.9. Thereby, we have generated a standard scalar for input data normalization and dimension reduction.

B. GMM CLUSTERING

The second step involves GMM-based clustering. Along the whole semiconductor manufacturing process flow, many process variations can affect the FT yield, causing yield distribution to show up in clusters with highly skewed or long tail trends at the lower yield range. The challenge in our prediction model framework is that the production process variation parameters are unknown. GMM can be effectively used to cluster Gaussian distributions based on the EM algorithm. It is assumed that yield variation would be relatively lower if the material has gone through a similar fabrication process flow. By clustering datasets based on FT yield, we can minimize the noise caused by other process variations and focus specifically on the WAT parameter analyses. GMM formulations with spherical, diagonal, full and tied covariance matrices and number of components from 1 to 6 are compared. The GMM model structure with the lowest BIC score is selected for the dataset clustering. In practice, there is no prior knowledge of which cluster a particular test data set would belong to. In order to avoid data leakage problem, the dataset is subject to GMM clustering only after the training and test datasets are split in our proposed flow.

C. FEATURE SELECTION AND MODEL SELECTION

Many of the previous attempts [6], [30], [31] to feature selection and dimension reduction tend to use all of the training and validation dataset before cross validation. However, this will result in data leakage problem because the validation dataset information is already included for feature selection. In real world application, we will only have the training dataset available at our disposal during feature and model selection.

In this section, we propose a nested 10-fold feature selection cross validation method to avoid the data leakage problem. The whole dataset is split into 90% training and validation dataset and 10% test dataset. The 90% dataset is used for the nested 10-fold feature selection and model selection cross validation step. A ten-fold Stratified ShuffleSplit cross validation method is then used instead of a regular cross validation as recommended by Kohavi *et al.* in Ref. [32]. At each fold, 60% of the dataset is used for training and 30% is used for validation. By using this method, we can randomly sample the dataset iteratively and simulate a practical situation wherein the batch of wafers is only partially tested, and we need to use incomplete production data for feature and model selection.

During each fold, the training dataset is used for feature selection and model evaluation. Two feature selection

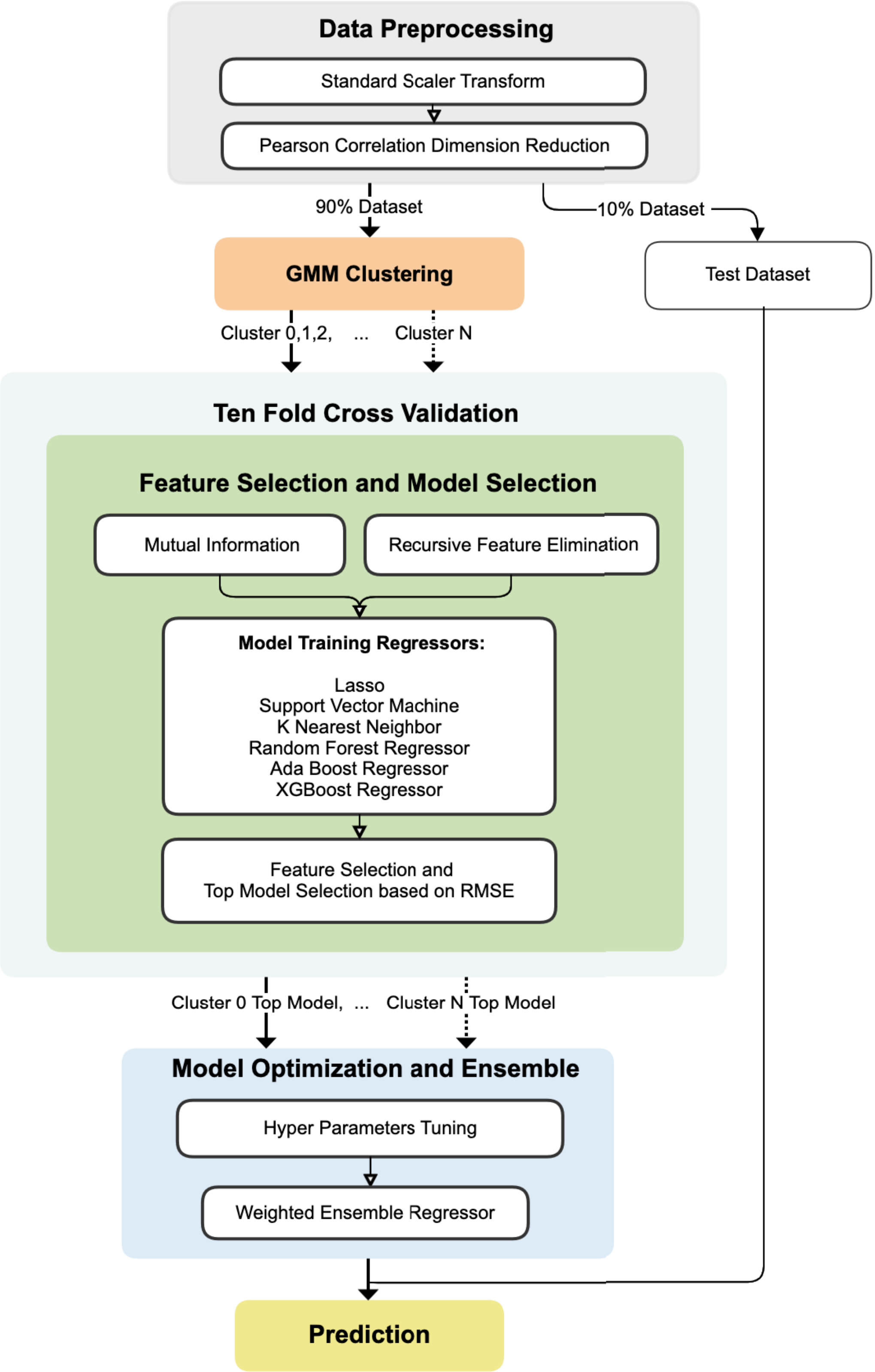


FIGURE 1. Overall computational flowchart procedure for Final Test yield prediction in this study.

methods are evaluated: MI and RFE OLS (defined earlier in Section II). Six popular and diversified regression models are selected for comparison. The Lasso regression is one type of linear regression model to solve high dimensional problems. It uses L1 regularization for feature selection which reduces model complexity and improves the prediction performance. Support vector machine (SVM) uses hyperplanes to maximize the separation between classes. Kernel tricks can be used to solve for both linear and non-linear problems. *K* Nearest Neighbor (KNN) is a simple and powerful algorithm. It classifies a new datapoint based on a similarity measure which is the distance function. The ease of interpretation and implementation of KNN makes the algorithm widely usable in pattern recognition and many other areas. The remaining regressors are decision tree based algorithms. Random Forest Regressor (RFR) is a reliable and efficient model using the bagging technique and aggregates multiple decision trees [33]. The Adaptive Boosting (ADA) technique, which is one of the popular boosting techniques, is often referred to as the best out-of-the-box [34] classifier when the decision tree is used as a base estimator. XGBoost (XGB) is also a decision-tree-based model using gradient boosting framework with good variance bias trade-off and fast execution speed.

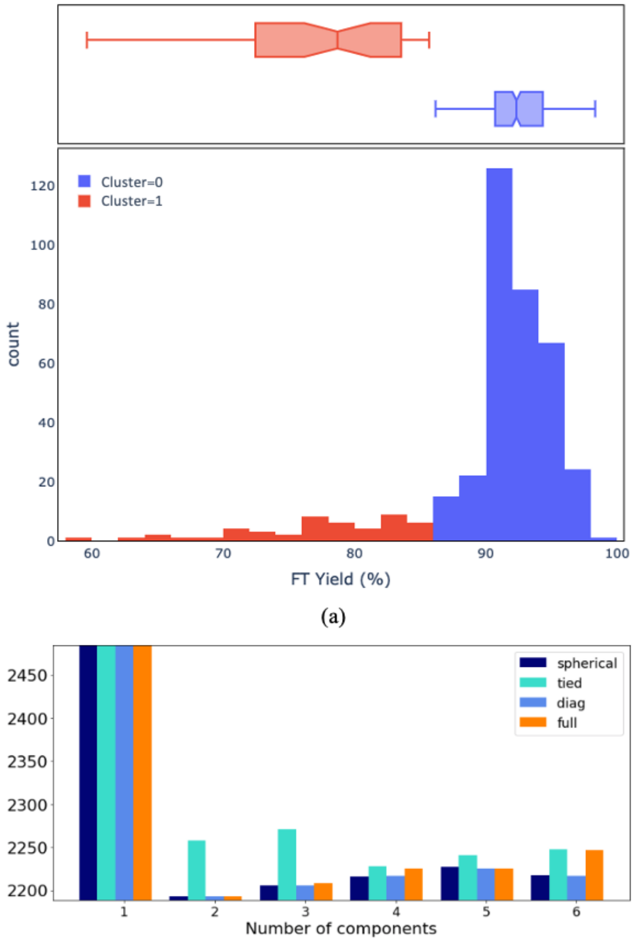
The metric for assessing model performance is taken to be the Root Mean Square Error (RMSE). At each fold, the model with the lowest RMSE is selected as the candidate model. A 10-fold averaged RMSE is used to compare and decide which feature selection method should be used. The mean and standard deviation of the RMSE value are taken into consideration for top model selection. Each cluster is running the cross validation independently. Therefore, the top candidate models can be different for each cluster.

D. MODEL OPTIMIZATION AND MODEL ENSEMBLE

After the top model is selected for each cluster, we use the 10-fold Stratified ShuffleSplit Grid Search [23] method for models' hyper parameters' optimization. The Grid Search score is set as the negative mean squared error and 90% of the dataset is used in this step. The next step is to build a weighted ensemble regressor using the top model from each cluster. All the top models' prediction results are combined with different weights to provide a final result. The weight is defined as the percentage of each cluster's data size over the 90% dataset (including both training and validation) size. Following this, the GMM clustering based ensemble regressor is generated. Finally, an unbiased test result is computed using the 10% test dataset, which is untouched from the beginning of the whole process.

V. RESULTS AND DISCUSSION

The recent three months of production data for two different new chip product lines - Device A and Device B, from Silicon Laboratories International are used here to test and validate our yield prediction regressor framework. We focus on Device A's model training procedure in this section as its FT yield distribution is more widely spread and has more low yield problems that would be worth analyzing. The procedure for data analysis pertaining to Device B is also the same and the results for it will be simply summarized in the form of a table at the end of this section.



(a)

(b)

FIGURE 2. **GMM clustering results for Device A. (a) FT yield distribution for cluster 0 and cluster 1, (b) BIC results comparison for different number of components and covariance matrices. Here, n = 2 and a diagonal covariance metric result in the lowest BIC score value.**

After data cleaning through removal of missing and invalid data, Device A production data set includes 432 backend lots with FT yield range all the way from 59.61% to 98.32%. Each backend lot consists of 1-2 wafers with the lot size ranging anywhere between 3000 to 10000 dies. Each wafer has 84 numeric WAT parameters as input data. For the backend lot with two wafers, the input data WAT measurements are averaged. After input data pre-processing, the number of input data parameters for Device A reduced from 84 to 61. After training and test dataset split step, the GMM analysis on the training dataset reveals that the data optimally comprises of two clusters (sub-distributions) as shown in Fig. 2(a) and that the best covariance option is diagonal, corresponding to the lowest BIC score, as shown in Fig. 2(b). Therefore, we will

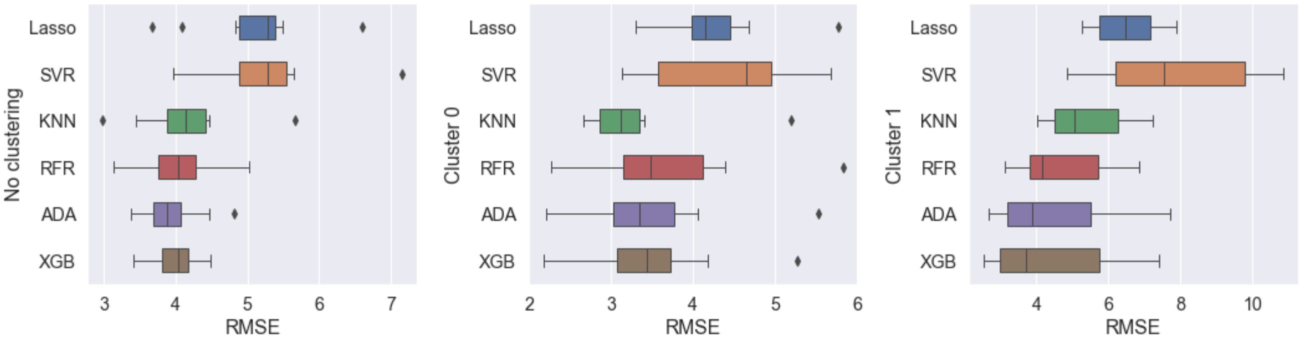


FIGURE 3. Boxplot for Device A 10-fold cross validation RMSE results with six different machine learning models. Left most plot is validation result without using GMM clustering. Middle and right plots are cluster 0 and cluster 1 results using GMM clustering method.

TABLE 1. Device A 10-fold cross validation RMSE results comparison between no-clustering and GMM clustering methods.

	No Clustering		Cluster 0		Cluster 1	
Models	*Mean*	*Stdev*	*Mean*	*Stdev*	*Mean*	*Stdev*
Lasso	5.101	0.802	4.227	0.708	6.498	0.938
SVR	5.270	0.843	4.370	0.852	7.830	2.238
KNN	4.152	0.706	**3.272**	0.725	5.366	1.142
RFR	4.040	0.501	3.712	0.973	4.711	1.341
ADA	**3.934**	0.454	3.491	0.885	4.433	1.794
XGB	4.087	0.350	3.562	0.782	**4.005**	1.289

need to do the training and validation analysis separately for these two clusters. Cluster 0 has 340 datapoints with a mean yield of 92.41% Cluster 1 (the lower extended tail of the yield distribution) has 48 datapoints with a mean yield of 79.29%. The yield distribution is right skewed and cluster 0 tends to be more normally distributed and cluster 1 has a wide scattered spread at the low yield range.

From the 10-fold cross validation results, the RFE OLS feature selection showed the lowest average RMSE of 3.651% compared to MI method with RMSE of 3.731%. Therefore, RFE OLS is selected for feature selection for subsequent model optimization and model ensemble learning. During the model training and selection step, the 10-fold cross validation RMSE mean and standard deviation are compared for the case of with and without GMM-based clustering. Detailed results are presented in Table 1 and in Fig. 3 as well. To compare the performance of our proposed GMM clustering ensemble regressor, an equal weighted ensemble regressor using a similar approach but without GMM clustering procedure is examined. From Fig. 3, we can see that the non-clustering based overall RMSE results are better than cluster 0 and cluster 1. This should come as no surprise since the size of the data set used for training here are more than that available for an individual cluster modeling and analysis alone.

For the non-clustering method, the tree based regressor outperforms the other models. The best performing models are ADA and XGB. These two models results are very close. Both of them have lower RMSE mean and standard deviation results compared to the other four models. The RFR RMSE mean is slightly lower than that of XGB but standard deviation is higher. Overall, the performance of XGB is better than that of RFR. For the clustering-based method, the best models are different for each cluster. For Cluster 0, it is KNN with an average RMSE of 3.272% and second lowest standard deviation of 0.725%. For Cluster 1, XGB is the best performing model with lowest average RMSE of 4.005% and relatively small standard deviation of 1.289% compared to the other models.

For the subsequent model optimization step, the number of input parameters are further reduced to 16 WAT parameters by using RFE OLS. After optimization, a weighted regressor is generated using KNN and XGB. The weight for cluster 0 is 340/388 i.e. 0.876 and weight for cluster 1 is 48/388 i.e. 0.124, based on the number of data points in each cluster. The cluster based regressor yields an RMSE of 2.221% for the final 10% test data set prediction which is a 46.9% reduction compared to no clustering regressor with a much higher RMSE of 4.183%. These results are stated in Table 3. Clearly, the use of clustering and subsequent ensemble regression accounting for the individual cluster contributions results in a far more improvised model learning and prediction outcome when compared to the standard cluster-free data analysis approach.

Coming to the analysis of Device B data, it includes 735 backend lots with the FT yield ranging from 66.44% to 98.67%. Each wafer has 120 numeric WAT parameters. After data pre-processing, the number of input parameters reduced from 120 to 73. In this case, the GMM clustering approach identified three discrete clusters. The 10-fold cross validation results are presented in Table 2 and the best performing models are highlighted in bold for each cluster. For the approach without clustering, RFR outperforms all the other five models and hence, the analysis from RFR is used for comparison with the split analysis based on the GMM ensemble regressor. The prediction involving the 10% test data set results are presented in Table 4. Once again, we see a significant error reduction of 16.3% in the RMSE value when the GMM-based cluster approach is applied. The test results are comparable

TABLE 2. Device B 10-fold cross validation RMSE results comparison between no-clustering and GMM clustering methods.

	No Clustering		Cluster 0		Cluster 1		Cluster 2	
Models	*Mean*	*Stdev*	*Mean*	*Stdev*	*Mean*	*Stdev*	*Mean*	*Stdev*
Lasso	6.834	0.320	3.035	0.185	2.656	0.129	6.790	1.444
SVR	8.045	0.297	3.096	0.181	2.658	0.102	4.754	0.570
KNN	6.922	0.306	3.253	0.164	2.780	0.172	**4.672**	0.502
RFR	**5.878**	0.242	**2.926**	0.083	2.644	0.113	5.002	0.694
ADA	6.165	0.235	3.063	0.108	**2.530**	0.084	5.121	0.632
XGB	6.078	0.419	3.264	0.164	2.629	0.162	4.921	0.898

TABLE 3. Device A test dataset RMSE results comparison between no clustering and GMM clustering method.

Method	Top Models	Test Dataset RMSE
GMM Clustering	KNN, XGB	2.221
No Clustering	ADA, XGB	4.183

TABLE 4. Device B test dataset RMSE results comparison between no clustering and GMM clustering method.

Method	Top Models	Test Dataset RMSE
GMM Clustering	RFR, ADA, KNN	3.240
No Clustering	RFR	3.872

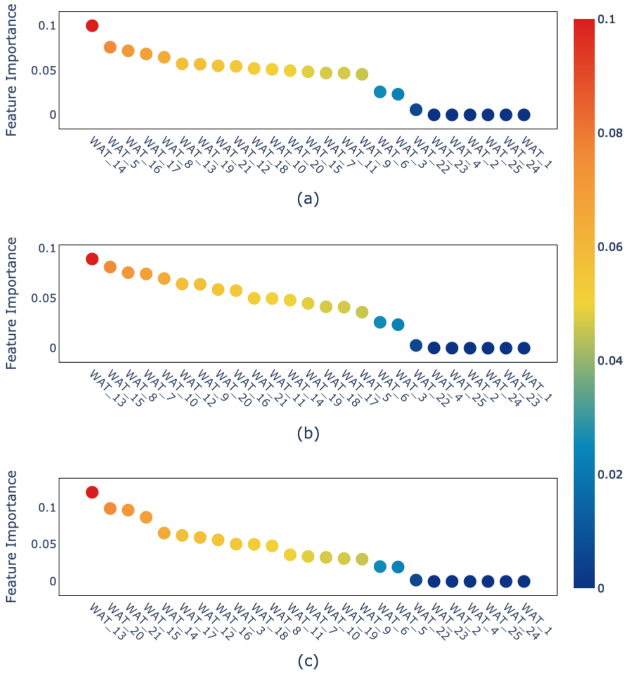


FIGURE 4. WAT parameters' feature importance analysis results for (a) Cluster-0 (b) Cluster-1 and (c) Cluster-2 showing the 25 most sensitive input parameters. There are significant differences in the order of importance of the WAT parameters for these three clusters, which highlights the importance of the GMM based yield clustering process in the overall machine learning framework.

with the absolute error prediction results for probe yield in Ref. [15]. Therefore, our proposed GMM ensemble regressor is effective and robust for FT yield prediction.

The hardware configuration on which the simulations were executed comprises of a personal laptop with 2.9GHz Intel Core i9 processor, 16GB 2400 MHz DDR4 RAM and Radeon Pro 560X GPU. The total training time for Device A is 6.44 mins, while the testing time is 3.03 sec. Similarly, the training time for Device B is 10.01 mins, while the testing time is 3.81 sec.

VI. FEATURE IMPORTANCE ANALYSIS AND YIELD IMPROVEMENT VALIDATION

Based on the above analysis, we have generated optimal models for the individual yield clusters. As a next step, we can adjust the most sensitive WAT parameters at the WF stage for FT yield improvement. In this section, we will discuss the feature importance analysis and examine the validity of our framework for yield improvement for one of the devices explored in this study, Device B, fabricated using 55 nm CMOS technology.

From Device B's GMM clustering results (which shows a trimodal yield pattern), the mean FT yield for Cluster-0, Cluster-1 and Cluster-2 are 92.92%, 82.75% and 70.81%. The top regression models for these three clusters turn out to be RFR, ADA and KNN, respectively. For decision tree based regressors like RFR and ADA, the Mean Decrease in Impurity (MDI) importance is the common method used for feature importance analysis. The MDI of a feature is computed using the weighted mean of the individual trees' improvement in the splitting criterion produced by each parameter, as described in Ref. [33]. Since the feature importance is not defined for the KNN and non-linear kernel SVR, we use RFR for the WAT parameter analysis of Cluster-2. The feature importance analysis ranking trends for the top 25 WAT parameters for each cluster, fitted with RFR for Cluster-0 and Cluster-2 and ADA model for Cluster-1, are plotted in Fig 4.

It can be seen that the most important WAT parameters are very different for each of the individual clusters. The top three WAT parameters for Cluster-0 are *WAT_14*, *WAT_5* and *WAT_16*. As for Cluster-1, they are *WAT_13*, *WAT_15* and *WAT_8* and for Cluster-2, it is *WAT_13*, *WAT_20* and *WAT_21*. Note that *WAT_13* is the common WAT parameter between Cluster-1 and Cluster-2 (the two lowest yield clusters which we intend to target for yield enhancement), which means that the drift in *WAT_13* could possibly be a major root cause for the low yield problem.

With our feedback on the top five sensitive WAT parameters for Cluster-1 and Cluster-2 to the foundry engineers, a dedicated DOE was formulated with various configurations of the parameter values by shifting their mean using sigma values (standard deviation of the parameter) calculated using historical production data. The details of the DOE configurations are listed in Table 5, wherein the positive and negative values are indicative of the increase or decrease in the mean value of the control parameter using a multiplicative factor to the sigma, where 0 obviously stands for no change whatsoever in the value of the parameter. The specific number of sigmas chosen for each WAT parameter shift is purely based on foundry engineer's experience. Values of mean adjustment

TABLE 5. DOE lots with top 5 features' adjustment based on each feature's sigma value and validation results showing the comparison between actual FT yield and predicted FT yield.

	Top 5 Features' Adjustment					FT Yield	
DOE Lots	*WAT_13 (sigma)*	*WAT_15 (sigma)*	*WAT_8 (sigma)*	*WAT_20 (sigma)*	*WAT_21 (sigma)*	Actual (%)	Predicted (%)
Lot#1	1.5	1.5	-1.5	1.5	1.5	97.2	87.5
Lot#2	1.5	0	0	0	0	96.7	86.8
Lot#3	0	1.5	0	1.5	1.5	89.6	88.1
Lot#4	0	0	-2	0	0	85.5	87.0
Lot#5	1.5	-1.5	1.5	-1.5	-1.5	84.7	87.1
Lot#6	0	0	2	0	0	77.0	72.5
Lot#7	0	-1.5	0	-1.5	-1.5	71.3	71.4
Lot#8(Ref)	0	0	0	0	0	76.4	74.3

ranging between 1-2 sigma is a common practise for DOE. Our analysis reveals that *WAT_15*, *WAT_20* and *WAT_21* are found to be moderately correlated WAT parameters during wafer manufacturing. As such, the DOE settings for these three parameters are kept in sync. These three parameters are moderately related to each other by a non-ignorable Pearson correlation between 0.8-0.9, which makes it necessary for them to be included in the analysis. As a result, during the feature selection step, they are not removed from the importance feature list based on the Pearson correlation results. *Lot*#3 and *Lot*#7 are designed to validate the yield change by increasing and decreasing these three WAT parameters. Similarly, *Lot*#4 and *Lot*#6 are designed to validate the effect of *WAT_8*. Based on foundry engineers' analysis, increasing *WAT_13* will help yield improvement. Therefore, *Lot*#2 is used to validate the effect to yield by only increasing *WAT_13*. *Lot*#1 and *Lot*#5 are used to evaluate the yield change with the combination of 5 WAT parameters' adjustments.

The wafer lots were custom fabricated based on the DOE configurations, and the lots were tested under the same FT production environment after probe and assembly. The results for the actual FT yield and predicted FT yield using the GMM ensemble regressor are listed in the last two columns of Table 5. *Lot*#8 with actual FT yield of 76.3% is the reference lot without any WAT adjustments. Based on the actual FT yield results for *Lot*#2 (where the most sensitive parameter, *WAT_13*, is the only one changed), it can be seen that the FT yield shows a drastic increase all the way up to 96.7% by adjusting this single parameter. The best performance lot is *Lot*#1 which includes a decrease in *WAT_8*, along with increase of *WAT_13* and the 3 remaining WAT parameters. The overall RMSE by taking the average of the 7 lots' RMSE values between the actual and predicted yield is 5.327%, which may be perceived to be high. However, note that the root cause for poor prediction performance for *Lot*#1 and *Lot*#2 (which are the major contributors to the inflated error) is attributable to the way the input parameter values are collectively represented here by taking the average of their values across 9 sites in each wafer. Based on our investigation, the low yield problem caused by *WAT_13* is mostly localized to the central site of the wafer, while the other 8 sites do not really pose the low yield problem. For the *WAT_13* DOE, additional process control was therefore needed to artificially increase the central site value so as to reduce the sigma across the different sites probed, which explains the anomaly here for *Lot*#1 and *Lot*#2. It is to be noted that our model currently lacks the capability to capture the location related low yield problem. However, on the whole, by adjusting the 5 important WAT parameters in the right directions from their prior means, we have successfully demonstrated a humongous 27.2% FT yield improvement, with reference to *Lot*#8 for the lowest yield sub-population. The results of the DOE and its strong qualitative agreement to the actual FT yield results clearly prove that our GMM ensemble regressor is a robust approach to cluster (or bin the yield data) and thereby enable customized low yield root cause analysis and subsequent yield optimization.

VII. CONCLUSION

In this paper, we proposed a novel procedure to predict semiconductor manufacturing backend FT yield using the front-end WAT parameters using a suite of machine learning techniques. The GMM was used for yield data clustering in our procedure prior to building an ensemble regressor. Our proposed procedure effectively addresses two common challenges faced by semiconductor fab process optimization and yield engineering teams → the high dimensional input parameter space and complexity in process variations that arise during different phases of the process chain. Real process data sets from two relatively immature chip product lines were used to test and verify the robustness and validity of our procedure and the results indeed clearly prove that the GMM clustering approach provides for a much more improved prediction model for future root cause analysis and effective process optimization.

One limitation in this study is that only two feature selection methods were evaluated, and they were not suitable for all the models applied during the model selection step. Further studies can be done on other feature selection algorithms. Proper ranking metrics for the feature importance method can also be explored. The WAT parameters tend to provide very limited information on the wafer front-end process. In the future, we can include analysis involving more inline process related input parameters, which are generated even earlier than the WAT parameters. Furthermore, in this work, we only used numerical wafer measurements as input data.

The role of categorical information such as fabrication and test equipment as well as product configuration data cannot be undermined and their influence on the observed and predicted yield also needs to be accounted for in our future efforts by finding methods to include categorical input variables in our analysis framework.

The limitation for GMM is that it is sensitive to the initial guesses of the parameter values and can get stuck at local minima. The model selection and optimization method can be further explored to reduce the training time and computational resource. Besides, there is still scope for low yield root cause analysis flow enhancement. The method not only needs to automatically identify the most sensitive WAT parameters, but also needs to fine tune and recommend the optimal WAT parameters' range for production process yield optimization. Our future work will include physics / logic / knowledge-informed neural networks incorporating the inverse design concept to make better decisions in estimating the sweet spot in the multi-dimensional process parameter space for achieving the highest process yield, ensuring reliability and robustness in the optimized process conditions.

REFERENCES

[1] C.-M. Fan, R.-S. Guo, S.-C. Chang, and C.-S. Wei, "SHEWMA: An end-of-line SPC scheme using wafer acceptance test data," *IEEE Trans. Semicond. Manuf.*, vol. 13, no. 3, pp. 344–358, Aug. 2000.

[2] S. Pampuri, A. Schirru, G. Fazio, and G. De Nicolao, "Multilevel lasso applied to virtual metrology in semiconductor manufacturing," in *Proc. IEEE Int. Conf. Autom. Sci. Eng.*, Aug. 2011, pp. 244–249.

[3] K. B. Lee, S. Cheon, and C. O. Kim, "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 2, pp. 135–142, May 2017.

[4] I. Kovacs, M. Topa, A. Buzo, and G. Pelz, "An accurate yield estimation approach for multivariate non-normal data in semiconductor quality analysis," in *Proc. 14th Int. Conf. Synth., Modeling, Anal. Simulation Methods Appl. Circuit Design (SMACD)*, Jun. 2017, pp. 1–4.

[5] T. Yuan, S. Z. Ramadan, and S. J. Bae, "Yield prediction for integrated circuits manufacturing through hierarchical Bayesian modeling of spatial defects," *IEEE Trans. Rel.*, vol. 60, no. 4, pp. 729–741, Dec. 2011.

[6] C.-F. Chien, C.-W. Liu, and S.-C. Chuang, "Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement," *Int. J. Prod. Res.*, vol. 55, no. 17, pp. 5095–5107, Sep. 2017.

[7] S.-J. Jang, J.-S. Kim, T.-W. Kim, H.-J. Lee, and S. Ko, "A wafer map yield prediction based on machine learning for productivity enhancement," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 4, pp. 400–407, Nov. 2019.

[8] S.-J. Jang, J.-H. Lee, T.-W. Kim, J.-S. Kim, H.-J. Lee, and J.-B. Lee, "A wafer map yield model based on deep learning for wafer productivity enhancement," in *Proc. 29th Annu. SEMI Adv. Semiconductor Manuf. Conf. (ASMC)*, Apr. 2018, pp. 29–34.

[9] 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)*, Aug. 2017, pp. 744–749.

[10] L. Wu and J. Zhang, "Fuzzy neural network based yield prediction model for semiconductor manufacturing system," *Int. J. Prod. Res.*, vol. 48, no. 11, pp. 3225–3243, Jun. 2010.

[11] 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*, Jun. 2013, pp. 363–368.

[12] 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.

[13] Y. Meidan, B. Lerner, G. Rabinowitz, and M. Hassoun, "Cycle-time key factor identification and prediction in semiconductor manufacturing using machine learning and data mining," *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 2, pp. 237–248, May 2011.

[14] S. Mahadevan and S. L. Shah, "Fault detection and diagnosis in process data using one-class support vector machines," *J. Process Control*, vol. 19, no. 10, pp. 1627–1639, Dec. 2009.

[15] 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.

[16] J. Wang, J. Zhang, and X. Wang, "A data driven cycle time prediction with feature selection in a semiconductor wafer fabrication system," *IEEE Trans. Semicond. Manuf.*, vol. 31, no. 1, pp. 173–182, Feb. 2018.

[17] J. Wang, P. Zheng, and J. Zhang, "Big data analytics for cycle time related feature selection in the semiconductor wafer fabrication system," *Comput. Ind. Eng.*, vol. 143, May 2020, Art. no. 106362.

[18] T. Chen, "A PCA-FBPN approach for job cycle time estimation in a wafer fabrication factory," *Int. J. Fuzzy Syst. Appl.*, vol. 2, no. 2, pp. 50–67, Apr. 2012.

[19] T. Chen, "Job cycle time estimation in a wafer fabrication factory with a bi-directional classifying fuzzy-neural approach," *Int. J. Adv. Manuf. Technol.*, vol. 56, nos. 9–12, pp. 1007–1018, Oct. 2011.

[20] T. Chen, "Embedding a back propagation network into fuzzy C-means for estimating job cycle time: Wafer fabrication as an example," *J. Ambient Intell. Humanized Comput.*, vol. 7, no. 6, pp. 789–800, Dec. 2016.

[21] R. Suganya and R. Shanthi, "Fuzzy C-means algorithm—A review," *Int. J. Sci. Res. Publications*, vol. 2, no. 11, p. 1, 2012.

[22] B. Thomas and M. Nashipudimath, "Comparative analysis of fuzzy clustering algorithms in data mining," *Int. J. Adv. Res. Comput. Sci. Electron. Eng.*, vol. 1, no. 7, p. 221, 2012.

[23] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[24] B. Lenz and B. Barak, "Data mining and support vector regression machine learning in semiconductor manufacturing to improve virtual metrology," in *Proc. 46th Hawaii Int. Conf. Syst. Sci.*, Jan. 2013, pp. 3447–3456.

[25] M. Saqlain, B. Jargalsaikhan, and J. Y. Lee, "A voting ensemble classifier for wafer map defect patterns identification in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 2, pp. 171–182, May 2019.

[26] 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.

[27] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Comput. Surv.*, vol. 45, no. 1, pp. 1–40, Nov. 2012.

[28] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.

[29] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 1998, pp. 645–648.

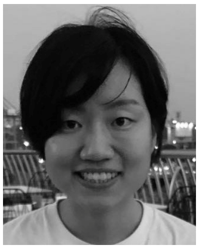
[30] P.-H. Chou, M.-J. Wu, and K.-K. Chen, "Integrating support vector machine and genetic algorithm to implement dynamic wafer quality prediction system," *Expert Syst. Appl.*, vol. 37, no. 6, pp. 4413–4424, Jun. 2010.

[31] 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.

[32] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. AI*, Montreal, QC, Canada, vol. 14, Aug. 1995, pp. 1137–1145.

[33] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[34] B. Kégl, "The return of AdaBoost.MH: Multi-class Hamming trees," 2013, *arXiv:1312.6086*. [Online]. Available: http://arxiv.org/abs/1312.6086



DAN JIANG received the B.Eng. degree in electrical and electronics engineering and the M.Eng. degree in communications engineering from Nanyang Technological University (NTU), Singapore, in 2013 and 2017, respectively. She is currently pursuing the Ph.D. degree with the Engineering Product Development (EPD) Pillar, Singapore University of Technology and Design (SUTD), under the Industry Postgraduate Program (IPP), supported by the Economic Development Board (EDB) of Singapore. Since 2013, she has been working as a Product Test Engineer with Silicon Laboratories International Pte. Ltd., on semiconductor product test and data analytics tools development. Her current research interests include semiconductor manufacturing yield prediction as well as big data and artificial intelligence for semiconductor process and quality optimization.

WEIHUA LIN received the B.Eng. degree in nuclear electronics from the University of Science and Technology of China (USTC), in 1991, and the M.Sc. degree in electronics engineering from the National University of Singapore (NUS), in 2001. He is currently the Senior Product Test Engineering (PTE) Director of Silicon Laboratories International Pte. Ltd. He has close to 30 years of working experience in several semiconductor companies, such as National Semiconductor, Lucent Microelectronics, and Silicon Labs. He has been focusing on integrated circuit (IC) test and product engineering as well as field application and customer support. He is also providing technical consultation roles in IC and module test development, product qualification, and yield optimization to achieve good product quality at lower possible cost.

NAGARAJAN RAGHAVAN (Member, IEEE) received the Ph.D. degree in microelectronics from the Division of Microelectronics, Nanyang Technological University (NTU), Singapore, in 2012. He is currently an Assistant Professor with the Engineering Product Development (EPD) Pillar, Singapore University of Technology and Design (SUTD). Prior to this, he was a Postdoctoral Fellow with the Massachusetts Institute of Technology (MIT) in Boston and with imec in Belgium, in joint association with Katholieke Universiteit Leuven (KUL). To date, he has authored/coauthored more than 200 international peer-reviewed publications and five invited book chapters as well. His research interests include reliability assessment, characterization, and lifetime prediction of nanoelectronic devices as well as material design for reliability, physics informed machine learning, uncertainty quantification, and prognostics and health management of electronic systems. He was an Invited Member of the IEEE GOLD Committee from 2012 to 2014. He was a recipient of the IEEE Electron Device Society (EDS) Early Career Award for 2016, the Asia-Pacific recipient for the IEEE EDS Ph.D. Student Fellowship in 2011, and the IEEE Reliability Society Graduate Scholarship Award in 2008. He also serves as the General Chair for IEEE IPFA 2021 at Singapore and has consistently served on the review committee for various IEEE journals and conferences, including IRPS, IIRW, IPFA, and ESREF. He is also serving as an Associate Editor for IEEE Access.