A Data-driven Approach to Material Removal Rate Prediction in Chemical Mechanical Polishing

Zhixiong Li[1] and Dazhong Wu[2]

[1,2]*Department of Mechanical and Aerospace Engineering, University of Central Florida, Orlando, FL 32816, USA*

zhixiong.li@Knights.ucf.edu
dazhong.wu@ucf.edu

ABSTRACT

Chemical mechanical polishing (CMP) has been widely used in the semiconductor sector for creating planar surfaces with the combination of chemical and mechanical forces. CMP is very complex because several chemical and mechanical phenomena (e.g., surface kinetics, contact mechanics, stress mechanics, and tribochemistry) are involved. Due to the complexity of the CMP process, it is very challenging to predict material removal rate (MRR) with sufficient accuracy. While physics-based methods have been introduced to predict MRR, little research has been reported on data-driven predictive modeling of MRR in the CMP process. This paper presents a novel decision tree-based ensemble learning algorithm that trains a predictive model of MRR on condition monitoring data. A stacking technique is used to combine three decision tree-based learning algorithms, including the random forests (RF), gradient boosting trees (GBT), and extremely randomized trees (ERT). The proposed method is demonstrated on the data collected from a wafer CMP tool that removes material from the surface of the wafer. Experimental results have shown that the decision tree-based ensemble learning algorithm can predict MRR in the CMP process with very high accuracy.

1. INTRODUCTION

Chemical mechanical polishing (CMP) was invented in IBM in the early 1980s to create a planar surface and enable subsequent lithographic imaging (Krishnan et al., 2009). The global CMP market in 2014 is valued at $3.32 billion and is estimated to reach $4.94 billion by 2020 (Marketsandmarkets.com). The key factors driving the growth of the CMP market are the increasing need of CMP for wafer polishing, high demand for consumer electronic products, and increasing use of micro-electro-mechanical systems. A typical CMP tool consists of a rotating table used to carry a polishing pad, a replaceable polishing pad attached to the table, a translating and rotating wafer carrier used to carry the wafer, a slurry dispenser, and a translating and rotating dresser used to condition the polishing pad (Steigerwald et al., 2008; Zantye et al, 2004). During the CMP process, a wafer is pressed against a polishing pad while a wafer carrier and a polishing pad are rotating in the same direction. An abrasive and corrosive chemical slurry is deposited onto the polishing pad during the CMP process. Modern CMP is a very complex process that involves several chemical and mechanical phenomena such as surface kinetics, electrochemical interfaces, contact mechanics, stress mechanics, hydrodynamics, and tribochemistry. The performance of the CMP process is measured using the metrics such as material removal rate (MRR), planarization (e.g., surface roughness), and process stability.

One of the key challenges in CMP is to achieve a high MRR and low non-uniformity of the polished surface (i.e., surface roughness). Fundamental understanding of the material removal mechanism in CMP is critical to control the CMP process and ultimately to control the quality of the polished surface. Because the performance of the CMP process is affected by many process variables such as the attributes of the slurry, polishing pad, and wafer carrier, it is very challenging to predict MRR with sufficient accuracy. The main contribution of this study is that a decision tree-based ensemble learning approach is introduced to predict MRR in CMP.

The remainder of this paper is organized as follows. Section 2 reviews the related work on CMP. Section 3 introduces an ensemble learning-based predictive modeling approach to MRR prediction. Section 4 presents a case study as well as discusses experimental results. Section 5 provides conclusions and future work.

2. RELATED WORK

Luo and Dornfeld (2001) proposed a physics-based model that predicts MRR by taking into account wafer hardness, pad hardness, pad roughness, abrasive size, and abrasive geometry. The experimental data collected from a silicon

CMP process were used to verify the model. The experimental results have shown that the predictive model can estimate MRR. Lin and Wu (2002) conducted a set of experiments to investigate the effects of polishing parameters such as rotating speed, applied pressure, flow rate of flurry on surface characteristics when polishing an aluminum-based rigid disk. The experimental results have shown that MRR increases as the applied pressure and relative velocity between the disk and the polishing pad increase. The experimental results also suggested that the Preston equation could be further modified to improve prediction accuracy. Lee and Jeong (2011) introduced a model that estimates MRR for copper CMP processes using a modified form of the Preston equation. A spatial parameter that is composed of three normalized parameters (i.e., normal contact stress, relative velocity between the wafer and the polishing pad, and chemical reaction rate) was added into the original Preston equation. Experimental results have shown that the modified Preston equation can be used to estimate MRR. Lee et al. (2013) proposed a semi-empirical MRR distribution model for SiO_2 CMP. This model incorporates the effects of the size, concentration, and distribution of particles as well as the slurry flow rate, polishing pad surface topography, material properties, and chemical reactions.

Kong et al. (2010) introduced a model-based method that integrates nonlinear Bayesian analysis and statistical modeling to estimate MRR. The particle filtering method was used for nonlinear Bayesian analysis to predict the CMP process state. A set of Cu-CMP experiments was conducted to collect vibration signals from a CMP machine. Experimental results have shown that the predictive model achieved a R^2 value of 0.96. Lih et al. (2008) introduced an approach to the prediction of MRR in Silicon CMP using an adaptive neuro-fuzzy inference system. Experimental results have shown that the predictive model trained by the ANFIS can achieve substantial improvements in prediction accuracy in comparison with neural networks and neuro-fuzzy modeling methods.

While previous research efforts have been focused on the development of physics-based and model-based predictive modeling techniques for CMP, few studies have been conducted to explore data-driven methods to predict MRR in CMP. To fill the research gap, an ensemble learning algorithm is introduced to predict MRR in this paper.

3. ENSEMBLE LEARNING

Ensemble learning is a data-driven method that combines multiple machine learning algorithms (also known as base learners) into one learning algorithm to improve the performance of predictive models (Džeroski et al., 2004; Zhou, 2012). The base learners can be aggregated to reduce a predictive variance by randomization ensemble (Evans et al., 2003), or to reduce a predictive bias by boosting ensemble (Friedman, 2001), or both by stacking ensemble. In general, the predictive model trained by ensemble learning outperforms that of individual base learners (Li et al., 2017). None of these ensemble methods outperforms other methods consistently. However, some empirical studies have shown that stacking outperforms boosting and randomization (Džeroski et al., 2004). Therefore, stacking is used to combine multiple base learners in this work. Stacking combines multiple classification or regression models using a meta-classifier or meta-regressor. The stacking technique includes two steps: training base learners and training a meta-algorithm. Fig. 1 illustrates the two-layer ensemble learning method using stacking.

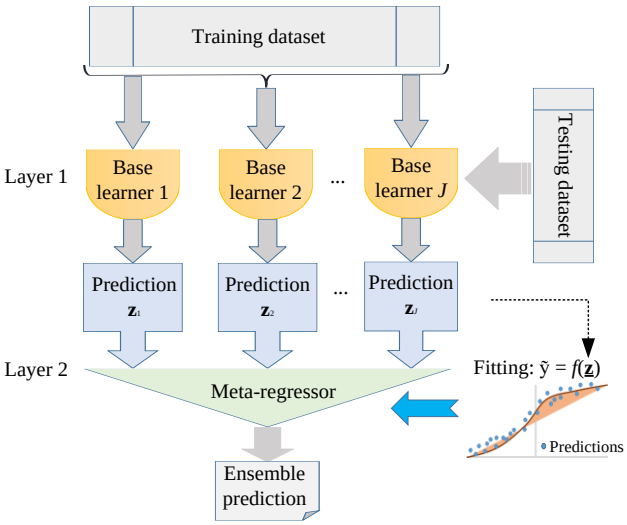

Figure 1. Two-layer ensemble learning using stacking

A training dataset is used to develop the predictive model. A validation or test dataset is used to validate the predictive model trained on the training dataset. The training, validation, and test datasets contain the raw sensory data. A set of features in the time and frequency domains is extracted from the training, validation, and test datasets. To reduce the dimensionality of the features, RF is used to reduce the number of features based on a measure called variable importance. In the model training phase, the selected features are fed into the decision tree-based ensemble learning algorithm that combines three base learning algorithms, including RF, GBT, and ERT. To develop a more accurate predictive model while avoiding overfitting, k-fold cross-validation (CV) is conducted to train the base learners. The predictions of the base learners are fed into another machine learning algorithm to train a meta-regressor. Two machine learning algorithms, including extreme learning machines (ELM) and classification and regression tree (CART) are used to train the meta-regressor. The output of the meta-regressor is the final prediction of the ensemble learning method. In the model validation phase, the validation and test datasets are used to validate the performance of the predictive model trained on the training dataset.

4. CASE STUDY

In this section, the decision tree-based ensemble learning method is demonstrated on the data acquired from the 2016 PHM Data Challenge (Propes and Rosca, 2016).

4.1 Data Description

The data contain multiple sensory signals collected from a wafer CMP tool that removes the material from the wafer surface. The total volume of the dataset is 187 MB. More details about the data can be found in Table 1. The conditions of four CMP tools were monitored during various runs of the CMP tools for specified wafers. The dataset is divided into three datasets, including one training dataset, one validation dataset, and one test dataset. The training data were collected from 1981 wafers under two operational stages: A and B. A total number of 672744 trajectories was collected from the wafers. These training trajectories are stored in 185 files in excel format. The validation and test datasets include 144148 and 156262 trajectories, respectively.

Table 1. Data description

ID	Description	ID	Description
1	Machine ID	14	Pressure applied to the retainer ring
2	Wafer ring location ID	15	Pressure applied to the ripple air bag
3	Time (s)	16	Usage of polishing membrane
4	Wafer ID	17	Usage of wafer carrier sheet
5	Stage ID (A or B)	18	Flow rate of slurry type A
6	Chamber ID	19	Flow rate of slurry type B
7	Usage of polish-pad backing film	20	Flow rate of slurry type C
8	Usage of dresser	21	Rotating rate of wafer
9	Usage of polishing table	22	Rotating rate of stage
10	Usage of dresser table	23	Rotating rate of head
11	Chamber pressure	24	Status of dressing water
12	Pressure applied to the main outer air bag	25	Pressure applied to the edge air bag
13	Pressure applied to the center air bag		

4.2 Feature Extraction and Selection

The raw data were transformed into a set of features and then a reduced subset of features before being processed by the decision tree-based ensemble learning algorithm. Four statistical features (see Eqs. 1-4) in the time domain, including the standard deviation, central moment, skewness and kurtosis, were extracted from each sensor signal. In addition, another three features in the frequency domain, including the maximum frequency amplitude, frequency center, and kurtosis of frequencies, were extracted from three vibration measurements. Eighty-five (85) features in total were extracted from the raw sensory data.

Standard deviation
$$\sigma(\mathbf{x}) = \mathrm{E}[\mathbf{x} - \mu]^{1/2} \quad (1)$$

Central moment
$$m_p(\mathbf{x}) = \mathrm{E}[\mathbf{x} - \mu]^p \quad (2)$$

Skewness
$$s(\mathbf{x}) = \mathrm{E}[\mathbf{x} - \mu]^3/\sigma^3 \quad (3)$$

Kurtosis
$$k(\mathbf{x}) = \mathrm{E}[\mathbf{x} - \mu]^4/\sigma^4 \quad (4)$$

where $\mathrm{E}[\cdot]$ denotes the expectation operation, μ is the mean value of $\mathbf{x}$, p is the order of moment and $p = 3$ in this study.

To avoid overfitting, a subset of the 85 features was selected in model training. RF was used to select the features by measuring the importance of features. More details about feature selection using RF can be found in Wu et al. (2018). The number of selected features was determined by balancing the trade-off between prediction accuracy and training time. Prediction accuracy is measured using R-square (R^2), root mean square error (RMSE), relative error (RE), and score function (S-score) (see Eqs. 5-8). RMSE and RE measure the deviations between the predicted and actual MRRs. R^2 measures the goodness of fit of a predictive model. The S-score, initially introduced in 2008 PHM Data Challenge, measures the performance of a model by taking into account whether the model overestimates and underestimates MRR.

RMSE
$$\varepsilon_{RMSE} = \sqrt{\mathrm{E}[(\hat{\mathbf{y}} - \mathbf{y})^2]} \quad (5)$$

RE
$$\varepsilon_{RPEi} = |\hat{y}_i - y_i| / y_i \quad (6)$$

S-score
$$\varepsilon_{CVi} = \begin{cases} \exp(-d_i/13), & d_i < 0 \\ \exp(d_i/10), & d_i \geq 0 \end{cases}, (d_i = \hat{y}_i - y_i) \quad (7)$$

R^2
$$\begin{cases} \varepsilon_{R^2} = 1 - SR/ST \\ SR = \sum_i \hat{y}_i - \bar{y}^T \\ ST = \sum_i \hat{y}_i - y_i^T \end{cases} \quad (8)$$

where $i = 1, 2, \ldots, N$ (N is the sample number), y_i is the actual MRR of the ith sample, $\hat{y}_i$ is the predicted MRR of the ith sample, $\hat{\mathbf{y}}$ is the matrix form of all predicted MRRs, and $\bar{y}$ is the mean value of the actual MRR vector $\mathbf{y}$.

To determine a subset of the initial features, GBT, RF, and ERT were used to train predictive models using 5, 20, 35, 50, 65, and 85 features. These decision tree-based learning algorithms were used to train predictive models using the training dataset. The validation dataset was used to evaluate the performance of the predictive models. Fig. 2 shows the average of R^2, RE, S-score, RMSE, and Training time. As shown in Fig. 2(a), R^2 increases as the number of features increases for both GBT and RF. R^2 decreases as the number of features exceeds 50 for ERT. As shown in Fig. 2(b), RE decreases as the number of features increases for both GBT and RF. RE increases as the number of features exceeds 35 for ERT. As shown in Fig. 2(c), S-score decreases as the

number of features increases for both GBT and RF. S-score increases as the number of features increases for ERT. As shown in Fig. 2(d), RMSE decreases as the number of features increases for both GBT and RF. RMSE increases as the number of features exceeds 35 for ERT. As shown in Fig. 2(e), training time does not vary with the number of features for both GBT and RF. However, training time increases as the number of features increases for ERT. Therefore, 35 out of 85 features were selected to train predictive models.

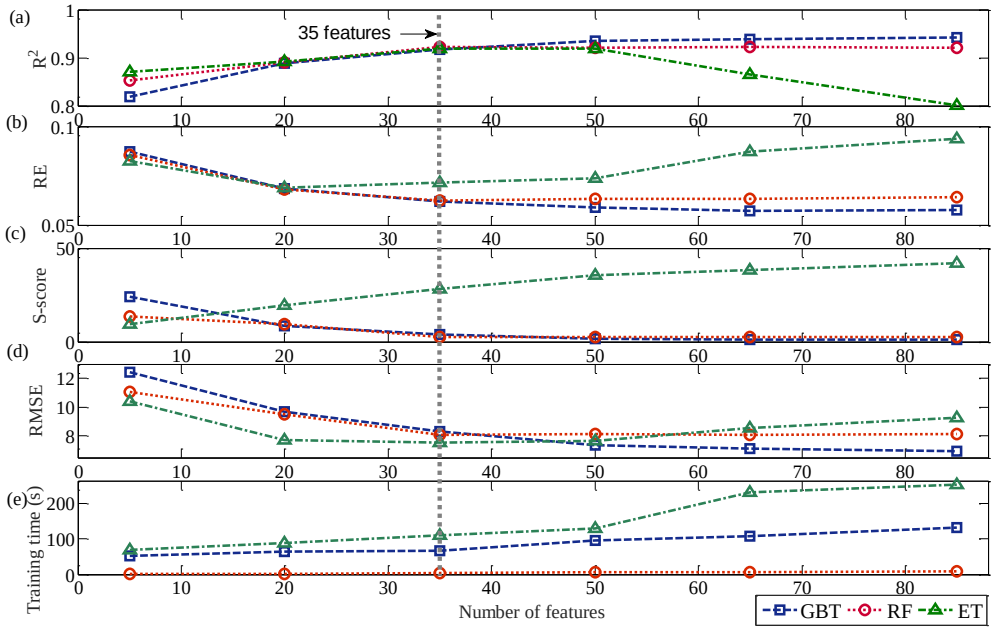

Figure 2. Prediction performance using different number of features: (a) R^2, (b) RE, (c) S-score, (d) RMSE, and (e) training time

4.3 Prediction Results Using Ensemble Learning

Table 2. Prediction performance

Validation Dataset				
Method	**R^2**	**RMSE**	**RE**	**S-score**
GBT	0.917	8.323	0.062	3.915
RF	0.917	8.066	0.063	2.476
ERT	0.919	7.571	0.064	5.521
CART-stacking	0.937	6.926	0.052	1.318
ELM-stacking	0.905	7.222	0.057	3.452
Test Dataset				
Method	**R^2**	**RMSE**	**RE**	**S-score**
GBT	0.919	8.252	0.058	2.224
RF	0.918	8.572	0.063	6.876
ERT	0.939	7.336	0.055	1.723
CART-stacking	0.941	7.009	0.056	1.034
ELM-stacking	0.94	7.261	0.054	2.051

The 35 features were fed into the decision tree-based ensemble learning algorithm. The predictive models trained by the ensemble learning methods were validated on the validation and test datasets. Table 2 lists the R^2, RE, S-score, RMSE values for CART-based stacking and ELM-based stacking methods. The experimental results have shown that the decision tree-based ensemble learning methods using CART and ELM as stacking methods outperform the base learners. For the validation dataset, the ensemble learning method using CART outperforms the ensemble learning method using ELM in terms of R^2, RE, S-score, RMSE. For the test dataset, the ensemble learning method using CART still outperforms the ensemble learning method using ELM in terms of R^2, S-score, and RMSE. However, the ensemble learning method using EML outperforms the ensemble learning method using CART slightly in terms of RE.

5. CONCLUSION

This paper has presented an ensemble learning-based prognostic approach to prediction of MRR in the CMP process. Two stacking techniques were used to combine RF, GBT, and ERT. This ensemble learning method was demonstrated on the datasets acquired from the 2016 PHM data challenge. The predictive model was developed on a training dataset, and then was validated on the validation and test datasets. The experimental results have shown that the decision tree-based ensemble learning approach predicts MRR of the CMP process with sufficient accuracy and

reasonable training time. In addition, the ensemble learning algorithm outperformed the base learners (i.e., RF, GBT, and ERT). In the future, the training process of the ensemble learning-based prognostics approach will be parallelized to improve the computation efficiency.

ACKNOWLEDGEMENT

The research reported in this paper is partially supported by the University of Central Florida (UCF) and the Digital Manufacturing and Design Innovation Institute (DMDII). Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the UCF and the DMDII.

REFERENCES

Džeroski, S., and Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one?. *Machine learning*, 54(3), pp. 255-273.

Evans, C., Paul, E., Dornfeld, D., Lucca, D., Byrne, G., Tricard, M., Klocke, F., Dambon, O., and Mullany, B. (2003). "Material removal mechanisms in lapping and polishing. *CIRP Annals-Manufacturing Technology*, 52(2), pp. 611-633.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189-1232.

Krishnan, M., Nalaskowski, J. W., and Cook, L. M. (2009). Chemical mechanical planarization: slurry chemistry, materials, and mechanisms. *Chemical reviews*, 110(1), pp. 178-204.

Kong, Z., Oztekin, A., Beyca, O. F., Phatak, U., Bukkapatnam, S. T., and Komanduri, R. (2010). Process performance prediction for chemical mechanical planarization (CMP) by integration of nonlinear Bayesian analysis and statistical modeling. *IEEE Transactions on Semiconductor Manufacturing*, 23(2), pp. 316-327.

Lee, H., and Jeong, H. (2011). A wafer-scale material removal rate profile model for copper chemical mechanical planarization. *International Journal of Machine Tools and Manufacture*, 51(5), pp. 395-403.

Lee, H., Jeong, H., and Dornfeld, D. (2013). Semi-empirical material removal rate distribution model for SiO 2 chemical mechanical polishing (CMP) processes. *Precision Engineering*, 37(2), pp. 483-490.

Li, Z., Wu, D., Hu, C., and Terpenny, J. (2017). An Ensemble Learning-based Prognostic Approach with Degradation-Dependent Weights for Remaining Useful Life Prediction. *Reliability Engineering & System Safety*.

Lih, W.-C., Bukkapatnam, S. T., Rao, P., Chandrasekharan, N., and Komanduri, R. (2008). Adaptive neuro-fuzzy inference system modeling of MRR and WIWNU in CMP process with sparse experimental data. *IEEE Transactions on Automation Science and Engineering*, 5(1), pp. 71-83.

Lin, S.-C., and Wu, M.-L. (2002). A study of the effects of polishing parameters on material removal rate and non-uniformity. *International Journal of Machine Tools and Manufacture*, 42(1), pp. 99-103.

Luo, J., and Dornfeld, D. A. (2001). Material removal mechanism in chemical mechanical polishing: theory and modeling. *IEEE transactions on Semiconductor Manufacturing*, 14(2), pp. 112-133.

Propes, N. and Rosca, J. (2016). PHM Society Data Challenge. https://www.phmsociety.org/events/conference/phm/16/data-challenge.

Steigerwald, J. M., Murarka, S. P., and Gutmann, R. J. (2008). Chemical mechanical planarization of microelectronic materials. John Wiley & Sons.

Wu, D., Jennings, C., Terpenny, J., Kumara, S., and Gao, R. X. (2018). Cloud-Based Parallel Machine Learning for Tool Wear Prediction. *Journal of Manufacturing Science and Engineering*, 140(4), 041005.

Zantye, P. B., Kumar, A., and Sikder, A. (2004). Chemical mechanical planarization for microelectronics applications. *Materials Science and Engineering*: R: Reports, 45(3), pp. 89-220.

Zhou, Z.-H. (2012). Ensemble methods: foundations and algorithms, CRC press.

BIOGRAPHIES

Dr. Zhixiong Li received his Ph.D. in Transportation Engineering from Wuhan University of Technology, China in 2013. He is a currently a Postdoctoral Fellow in the Department of Mechanical and Aerospace Engineering at the University of Central Florida. His research interests include mechanical system modeling and control, prognostics and health management, and system dynamics & condition monitoring. He serves as an associate editor for the *IEEE Access*.

Dr. Dazhong Wu received his B.S. from Hunan University, M.S. from Shanghai Jiao Tong University in China, and Ph.D. from the Georgia Institute of Technology, all in Mechanical Engineering. He is currently an Assistant Professor in the Department of Mechanical and Aerospace Engineering at the University of Central Florida. His research is focused on process monitoring, diagnostics, prognostics, smart manufacturing, reliability engineering, and big data analytics. He serves on editorial boards of Journal of Manufacturing Systems and Journal of Intelligent Manufacturing.