



Material removal rate prediction in chemical mechanical planarization with conditional probabilistic autoencoder and stacking ensemble learning

Yupeng Wei¹ · Dazhong Wu²

Received: 3 February 2022 / Accepted: 4 October 2022 / Published online: 14 October 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Chemical mechanical planarization (CMP) is a complex and high-accuracy polishing process that creates a smooth and planar material surface. One of the key challenges of CMP is to predict the material removal rate (MRR) accurately. With the development of artificial intelligence techniques, numerous data-driven models have been developed to predict the MRR in the CMP process. However, these methods are not capable of considering surface topography in MRR predictions because it is difficult to observe and measure the surface topography. To address this issue, we propose a graphical model and a conditional variational autoencoder to extract the features of surface topography in the CMP process. Moreover, process variables and the extracted features of surface topography are fed into an ensemble learning-based predictive model to predict the MRR. Experimental results have shown that the proposed method can predict the MRR accurately with a root mean squared error of 6.12 nm/min, and it outperforms physics-based machine learning and data-driven methods.

Keywords Chemical mechanical planarization · Deep learning · Graphical model · Material removal rate prediction

Introduction

Chemical mechanical planarization (CMP) refers to a high-precision surface polishing process with a combination of chemical and mechanical forces (Sheu et al., 2012; Zantye et al., 2004). CMP was initially innovated by Klaus D. Beyer in the 1980s to create a smooth surface so that lithographic imaging can be implemented subsequently (Krishnan et al., 2010). CMP can be used to polish a wide range of materials, such as tungsten, semiconductors, metal, carbon nanotubes, and silicon oxide (Awano, 2006; Steigerwald et al., 1997). CMP has been used in many applications, such as optical components, wireless communications, and large-scale integration manufacturing (Lee et al., 2016; Leon et al., 2017;

Yin et al., 2019). A typical CMP device includes a rotating table, a planarization pad, a wafer carrier, a wafer, a slurry dispenser, and a rotating dresser, where a wafer is captured by a wafer carrier, and a polishing pad is attached to the rotating table. In the CMP process, a wafer is pushed toward the planarization pad, and both the rotating table and the wafer carrier are rotated in an identical direction. The abrasive materials are dispensed on the planarization pad via a slurry during the polishing process. A rotating dresser may be engaged in conditioning the polishing pad after the CMP process.

The performance of the CMP process can be evaluated using many metrics, such as wafer-to-wafer thickness variation, surface roughness, and process reliability and stability. To reduce the wafer-to-wafer thickness variation in CMP, accurate prediction of material removal rate (MRR) is critical (Deng et al., 2021). However, predicting MRR with high accuracy remains a challenge because MRR depends on various process variables and surface topography, such as the rotating rate of the wafer, flow rate of slurry, polishing pad asperity density, wafer hardness, and so on (Park et al., 2008; Yu et al., 2016). According to the literature, numerous methodologies have been developed to predict the

✉ Yupeng Wei
yupeng.wei@sjsu.edu
Dazhong Wu
dazhong.wu@ucf.edu

¹ Department of Industrial and Systems Engineering, San Jose State University, San Jose 95192, CA, USA

² Department of Mechanical and Aerospace Engineering, University of Central Florida, Orlando 32816, FL, USA

MRR during the CMP process, and these methodologies can be classified into two groups: model-based and data-driven methods. The majority of model-based methods are built upon the basic or modified Preston equations (Luo et al., 1998). The Preston equation is an empirical model that considers the pressure applied to a wafer in a vertical direction and the relative speed between the wafer and the polishing pad (Evans et al., 2003). However, few model-based methods are able to accurately predict the MRR of the CMP process (Kong et al., 2010). Over the past few years, data-driven methods have been increasingly used to predict MRR by incorporating multiple process variables, such as rotating rate of wafer and flow rate of slurry (Lee & Kim, 2020; Xia et al., 2021). However, these methods are not capable of considering surface topography in MRR predictions as the surface topography is difficult to observe and measure (Chen et al., 2020). To address this issue, our contributions are listed as follows:

- A directed graphical model is proposed to reveal the relations among process variables, surface topography, and MRR during the CMP process.
- A conditional variational autoencoder is introduced based on the proposed directed graphical model to extract the features of the surface topography.
- An ensemble learning-based predictive model is developed to predict the MRR during the CMP process.

The remainder of this paper is organized as follows. Section [Related work](#) reviews the model-based and data-driven methods for MRR predictions in the CMP process. Section [Methodology](#) proposes a directed graphical model and introduces a conditional variational autoencoder to extract the features of the surface topography. In addition, an ensemble learning-based approach is presented in this section to predict the MRR during the planarization process. Section [Case study](#) uses a CMP dataset to demonstrate the effectiveness of the proposed method. Section [Conclusions and future work](#) concludes this study and directs future work.

Related work

This section reviews the model-based and data-driven methods for predicting MRR in CMP processes. The limitation of these methods is summarized at the end of this section.

Model-based methods

Model-based methods refer to the methods that predict the behavior of a system or a process using numerical or analytical models. Luo and Dornfeld (2001) presented a physics-based model to predict the MRR in the CMP process,

where both wafer-abrasive and pad-abrasive mechanisms in plastic contact mode were investigated. The proposed model considered multiple process variables in MRR predictions, such as pressure, velocity, pad roughness, and so on. The experimental results have shown that the proposed model enables an accurate MRR prediction and a better understanding of the abrasive mechanism in the CMP process. Lee and Jeong (2011) presented a semi-empirical CMP model to predict the MRR during the copper CMP process by combining the basic form of the Preston equation and a spatial parameter. The distributions of velocity, contact stress, and rate of reaction were considered in the proposed model. Zhao and Chang (2002) presented a closed-form equation to predict the MRR in the polishing process of silicon wafers based upon a micro-contact and wear model. The proposed equation incorporated multiple process variables, material parameters, and chemical parameters. Experimental results have suggested that the MRR is sensitive to wafer hardness, slurry type, and rotating speed. Oh and Seok (2009) combined a mechanical abrasive model with a slurry dispensation model to estimate the MRR for silicon dioxide in the CMP process. The effects led by both mechanical and chemical actions were included in MRR predictions. The experimental results have demonstrated that the proposed method can deal with the non-Prestonian behavior during the planarization process. Lee et al. (2013) introduced a MRR distribution model to predict the MRR in the planarization process. To estimate the parameters of the proposed model, a CMP experiment was conducted on different types of slurries. Nguyen et al. (2015) introduced a MRR analytical model by considering both the contact time of the planarization process and the kinematic mechanism. The numerical results have demonstrated that the non-conformity of the pad wear is due to the inconsistencies in both cutting path density and contact time.

Data-driven methods

Data-driven methods refer to the methods that guide decision making using data instead of physical models representing the behavior behind a system or a process. Kong et al. (2010) integrated a statistical learning model with a non-linear Bayesian method to predict the MRR of the CMP process. The particle filter was implemented to estimate the state of the CMP process, and vibration signals were used to predict the MRR. The numerical results have demonstrated that this approach can effectively predict the MRR during the planarization process. Li et al. (2019) presented an ensemble learning method to predict the MRR in the planarization process. Temporal and frequency-domain features were extracted from multiple sensor measurements and fed into the ensemble learning method. The numerical results have demonstrated that the proposed methodology can predict the MRR at different polishing stages with high accuracy.

Yu et al. (2019) introduced a physics-constrained machine learning method to predict the MRR. The Greenwood and Williamson contact model (Greenwood & Williamson, 1966; Johnson & Johnson, 1987) served as a predictive model to estimate the MRR, and the random forests method was used to estimate the topography terms in the Greenwood and Williamson contact model. Wang et al. (2017) used a deep neural network to predict the MRR during the planarization process based on the polishing process variables. The particle swarm optimization method was implemented to study the effect of the learning rate on prediction accuracy. The numerical results have demonstrated that the proposed deep learning approach can accurately predict the MRR under different operating conditions. Jia et al. (2018) introduced an adaptive polynomial neural network to predict the MRR. The features and predictive models were selected automatically, and two novel categories of features were introduced to improve the prediction performance.

In summary, numerous model-based and data-driven methods have been introduced to predict the MRR in the CMP process. However, most model-based methods are not able to predict the MRR with high accuracy due to the complexity of the CMP process. Some of the existing data-driven methods are effective in predicting the MRR, however, few data-driven methods predict the MRR by taking into account the surface topography information because it is difficult to measure the surface topography in the CMP process. To address these issues, the objective of this study is to develop a directed graphical model and a conditional variational autoencoder to extract the features of the surface topography. In addition, an ensemble learning-based predictive model is presented to predict the MRR during the CMP process.

Methodology

The proposed methodology includes three primary steps. First, a directed graphical model is proposed to reveal the relations among process variables, surface topography, and MRR in the CMP process. Second, a conditional variational autoencoder is introduced based on the proposed directed graphical model to extract the features of the surface topography. Third, both process variables and the extracted features of the surface topography are fed into an ensemble learning-based predictive model to predict the MRR in the CMP process. More details of these three steps are introduced in the following subsections.

Directed graphical model

A directed graphical model refers to a probabilistic model where the dependency of multiple variables is revealed in a directed graph (Airoldi, 2007). Figure 1 shows the proposed

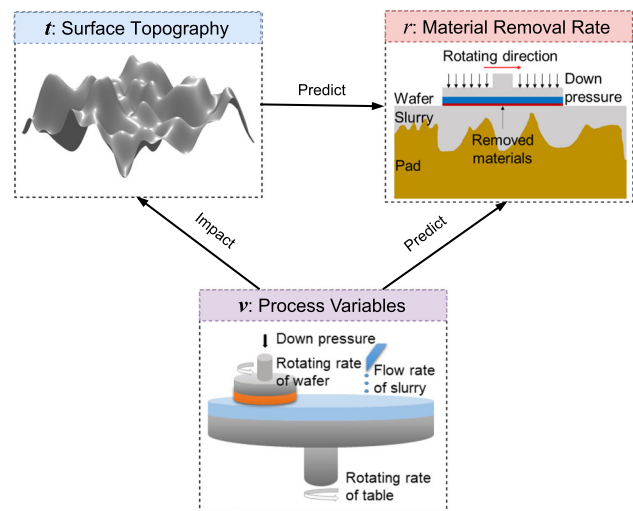


Fig. 1 The proposed directed graphical model, where **t** refers to the surface topography, **r** refers to the material removal rate, and **v** refers to the process variables

directed graphical model where the relationships among process variables, surface topography, and MRR are revealed.

The process variables of the CMP process, such as polishing pressure and flow rate of slurry, affect surface topography, such as pad asperity density and average asperity radii (Yu et al., 2019). Therefore, an arrow is considered to be directed from process variables to surface topography to represent that the process variables affect the surface topography. The process variables affect the MRR, for example, a higher rotation rate of wafer and table can lead to a higher MRR. Therefore, an arrow is directed from process variables to material removal rate to represent that the process variables can be used to predict the MRR. In addition, the surface topography can also affect the MRR, for instance, a higher active asperity density can lead to a higher MRR. Thus, a directed arrow is pointed from surface topography to material removal rate to represent that the surface topography can also be used to predict the MRR.

In the proposed graphical model, the process variables can be observed from sensor measurements and the MRR can be measured after the CMP process. However, the surface topography is difficult to observe and measure due to its dynamic evolution during the planarization process. To extract features that enable a maximized MRR prediction accuracy, the extracted features of the surface topography can be expressed as Eq. (1),

$$\mathbf{t} := \underset{\mathbf{t}}{\operatorname{argmax}} p_{\theta}(r | \mathbf{v}) \quad (1)$$

where **t** refers to the features of the surface topography, **r** is the material removal rate, **v** refers to the process variables, θ is a collection of parameters in the conditional probability of

MRR, i.e. $p_\theta(r | \mathbf{v})$. To simplify the optimization process, we use log-likelihood instead of likelihood. Then, Eq. (1) can be rewritten as:

$$\mathbf{t} := \underset{\mathbf{t}}{\operatorname{argmax}} \log p_\theta(r | \mathbf{v}) \quad (2)$$

With using the Bayesian theory, $\log p_\theta(r | \mathbf{v})$ can also be written as Eq. (3).

$$\begin{aligned} \log p_\theta(r | \mathbf{v}) &= \log \frac{p_\theta(\mathbf{v}, r)}{p_\theta(\mathbf{v})} \\ &= \log \frac{p_\theta(\mathbf{v}, r, \mathbf{t})}{p_\theta(\mathbf{t} | \mathbf{v}, r) \cdot p_\theta(\mathbf{v})} \end{aligned} \quad (3)$$

Based on the chain rule of the proposed directed graphical model, $p_\theta(\mathbf{v}, r, \mathbf{t})$ can be expressed as:

$$p_\theta(\mathbf{v}, r, \mathbf{t}) = p_\theta(r | \mathbf{t}, \mathbf{v}) \cdot p_\theta(\mathbf{t} | \mathbf{v}) \cdot p_\theta(\mathbf{v}) \quad (4)$$

By substituting Eq. (4) to Eq. (3), $\log p_\theta(r | \mathbf{v})$ can be written as Eq. (5).

$$\log p_\theta(r | \mathbf{v}) = \log \frac{p_\theta(r | \mathbf{t}, \mathbf{v}) \cdot p_\theta(\mathbf{t} | \mathbf{v})}{p_\theta(\mathbf{t} | \mathbf{v}, r)} \quad (5)$$

The conditional probability distribution of \mathbf{t} is unknown as the surface topography can not be obtained directly. Thus, $p_\theta(\mathbf{t} | \mathbf{v}, r)$ is intractable. To deal with this intractable posterior distribution, a variational inference is introduced and $\log p_\theta(r | \mathbf{v})$ can be expressed as:

$$\begin{aligned} \log p_\theta(r | \mathbf{v}) &= \log \frac{p_\theta(r | \mathbf{t}, \mathbf{v}) \cdot p_\theta(\mathbf{t} | \mathbf{v})}{p_\theta(\mathbf{t} | \mathbf{v}, r)} \cdot \frac{q_\phi(\mathbf{t} | \mathbf{v}, r)}{q_\phi(\mathbf{t} | \mathbf{v}, r)} \end{aligned} \quad (6)$$

Then, the expectation of Eq. (6) can be written as Eq. (7), where ϕ is the collection of parameters in the variational inference $q_\phi(\mathbf{t} | \mathbf{v}, r)$.

$$\mathbb{E}_{\mathbf{t} \sim q_\phi(\mathbf{t} | \mathbf{v}, r)} \left[\log \frac{p_\theta(r | \mathbf{t}, \mathbf{v}) \cdot p_\theta(\mathbf{t} | \mathbf{v})}{p_\theta(\mathbf{t} | \mathbf{v}, r)} \cdot \frac{q_\phi(\mathbf{t} | \mathbf{v}, r)}{q_\phi(\mathbf{t} | \mathbf{v}, r)} \right] \quad (7)$$

Equation (7) can be decomposed into the sum of two terms, where the first term can be expressed as:

$$\mathbb{E}_{\mathbf{t} \sim q_\phi(\mathbf{t} | \mathbf{v}, r)} \left[\log \frac{p_\theta(r | \mathbf{t}, \mathbf{v}) \cdot p_\theta(\mathbf{t} | \mathbf{v})}{q_\phi(\mathbf{t} | \mathbf{v}, r)} \right] \quad (8)$$

The second term is expressed as Eq. (9), which is a KL-divergence of two distributions.

$$D_{KL}[q_\phi(\mathbf{t} | \mathbf{v}, r) \| p_\theta(\mathbf{t} | \mathbf{v}, r)] \quad (9)$$

Because the KL-divergence of two distributions is always positive, and this KL-divergence term includes an intractable probability distribution $p_\theta(\mathbf{t} | \mathbf{v}, r)$. A variational lower bound is introduced, and the extracted features of surface topography can be expressed as:

$$\mathbf{t} := \underset{\mathbf{t}}{\operatorname{argmax}} \mathbb{E}_{\mathbf{t} \sim q_\phi(\mathbf{t} | \mathbf{v}, r)} \left[\log \frac{p_\theta(r | \mathbf{t}, \mathbf{v}) \cdot p_\theta(\mathbf{t} | \mathbf{v})}{q_\phi(\mathbf{t} | \mathbf{v}, r)} \right] \quad (10)$$

The expectation term of Eq. (10) can also be decomposed into two terms, and the extracted features of the surface topography can be written as:

$$\begin{aligned} \mathbf{t} := \underset{\mathbf{t}}{\operatorname{argmax}} \mathbb{E}_{\mathbf{t} \sim q_\phi(\mathbf{t} | \mathbf{v}, r)} &[\log p_\theta(r | \mathbf{t}, \mathbf{v}) \\ &- D_{KL}[q_\phi(\mathbf{t} | \mathbf{v}, r) \| p_\theta(\mathbf{t} | \mathbf{v})] \end{aligned} \quad (11)$$

Based on the Universal Approximation Theorem (Hornik et al., 1989), neural networks are employed to approximate the three conditional probability distributions in Eq. (11).

Conditional probabilistic autoencoders

The conditional probability distributions in the variational lower bound are approximated using autoencoder-based neural networks. The conditional probability $q_\phi(\mathbf{t} | \mathbf{v}, r)$ is approximated with an encoder network. The inputs of this encoder network are process variables \mathbf{v} and the MRR r , and the outputs of this network are the features of the surface topography \mathbf{t} . We name this encoder network the generative encoder network as it aims at generating the features of the surface topography. The relationships between the inputs and the outputs of this network can be mathematically written as Eq. (12),

$$\begin{aligned} \mathbf{o}_{q,L} &= f_{q,L} \cdots [f_{q,l} \cdots [f_{q,2}[f_{q,1}(\mathbf{v}, r)]]] \\ \boldsymbol{\mu}_1 &= \mathbf{w}_{q,L+1,1} \cdot \mathbf{o}_{q,L} + \mathbf{b}_{q,L+1,1} \\ \operatorname{diag}(\boldsymbol{\Sigma}_1) &= \mathbf{w}_{q,L+1,2} \cdot \mathbf{o}_{q,L} + \mathbf{b}_{q,L+1,2} \\ \mathbf{t}_1 &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_1 \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{aligned} \quad (12)$$

where $f_{q,l}(\cdot)$ can be expressed as $f_{q,l}(\cdot) = \sigma(\mathbf{w}_{q,l} \cdot \mathbf{o}_{q,l-1} + \mathbf{b}_{q,l})$; $\mathbf{w}_{q,l}$ refers to the vector of weights of the generative encoder network at hidden layer l ; $\mathbf{b}_{q,l}$ is the bias vector of the generative encoder network at hidden layer l ; $\mathbf{o}_{q,l-1}$ is the output of the hidden layer $l-1$; σ refers to the activation function; $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ are mean and standard deviation of the conditional probability distribution $q_\phi(\mathbf{t} | \mathbf{v}, r)$ respectively.

The conditional probability $p_\theta(\mathbf{t} | \mathbf{v})$ is approximated with an encoder network. The inputs of this network are process variables, and the outputs are the features of the surface topography. We name this network as the conditional prior network as it aims at generating the features of the surface topography conditioning on the prior knowledge of process

variables. The relationships between the inputs and the outputs of this network can be mathematically written as Eq. (13),

$$\begin{aligned} \mathbf{o}'_{p,L} &= f'_{p,L} \cdots [f'_{p,l} \cdots [f'_{p,2}[f'_{p,1}(\mathbf{v})]]] \\ \boldsymbol{\mu}_2 &= \mathbf{w}'_{p,L+1,1} \cdot \mathbf{o}'_{p,L} + \mathbf{b}'_{p,L+1,1} \\ \text{diag}(\boldsymbol{\Sigma}_2) &= \mathbf{w}'_{p,L+1,2} \cdot \mathbf{o}'_{p,L} + \mathbf{b}'_{p,L+1,2} \\ \mathbf{t}_2 &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_2 \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{aligned} \tag{13}$$

where $f'_{q,l}(\cdot)$ can be written as $f'_{q,l}(\cdot) = \sigma(\mathbf{w}'_{q,l} \cdot \mathbf{o}'_{q,l-1} + \mathbf{b}'_{q,l})$; $\mathbf{w}'_{q,l}$ refers to the vector of weights of the conditional prior encoder network at hidden layer l ; $\mathbf{b}'_{q,l}$ is the bias vector of the conditional prior encoder network at hidden layer l ; $\mathbf{o}'_{q,l-1}$ is the output of the hidden layer $l - 1$; $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ are mean and standard deviation of the conditional probability distribution $q_\phi(\mathbf{t} | \mathbf{v})$ respectively; and L refers to the total number of hidden layers.

The conditional probability $p_\theta(r | \mathbf{t}, \mathbf{v})$ is approximated with a decoder network. The inputs of this network are process variables and the features of the surface topography extracted from the generative encoder network, and the outputs of this network are predicted MRR. We name this network as the predictive network as it aims at predicting the MRR in the CMP process. The relationships between the inputs and the outputs of this network can be mathematically expressed as Eq. (14),

$$\hat{r} = f_{p,L} \cdots [f_{p,l} \cdots [f_{p,2}[f_{p,1}(\mathbf{v}, \mathbf{t}_1)]]] \tag{14}$$

where $f_{p,l}(\cdot)$ can be written as $f_{p,l}(\cdot) = \sigma(\mathbf{w}_{p,l} \cdot \mathbf{o}_{p,l-1} + \mathbf{b}_{p,l})$; $\mathbf{w}_{p,l}$ refers to the vector of weights of the predictive decoder network at hidden layer l ; $\mathbf{b}_{p,l}$ is the bias vector of the predictive decoder network at hidden layer l ; $\mathbf{o}_{p,l-1}$ is the output of the hidden layer $l - 1$;

Then, the expectation of the variational lower bound in Eq. (11) can be considered as the MRR prediction errors, which can be rewritten as Eq. (15),

$$\mathbb{E}_{\mathbf{t} \sim q_\phi(\mathbf{t}|\mathbf{v},r)}[\log p_\theta(r | \mathbf{t}, \mathbf{v})] = \|r - \hat{r}\|_2^2 \tag{15}$$

where r is the ground truth of the MRR, and \hat{r} refers to the predicted MRR. The KL-divergence of the variational lower bound in Eq. (11) can be considered as the differences between two distributions, which can be expressed as Eq. (16).

$$\begin{aligned} D_{KL}[q_\phi(\mathbf{t} | \mathbf{v}, r) \| p_\theta(\mathbf{t} | \mathbf{v})] &= \frac{1}{2} \left(\text{tr}(\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1) \right. \\ &\left. + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - d + \log \left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \right) \right) \end{aligned} \tag{16}$$

Next, the gradient descent method can be used to train the parameters in these networks. However, it may not be optimized to generate \mathbf{t}_2 as the conditional prior network is not connected to a predictive network. To address this issue, another predictive network is introduced. A similar setup can also be found in Zhao et al. (2017), Pandey and Dukkupati (2017), and Wei et al. (2021). The inputs of this network are process variables and the features of the surface topography extracted from the conditional prior encoder network, and the outputs of this predictive network are predicted MRR. The relationships between the inputs and the outputs of this predictive network can be mathematically expressed as Eq. (17),

$$\hat{r}' = f''_{p,L} \cdots [f''_{p,l} \cdots [f''_{p,2}[f''_{p,1}(\mathbf{v}, \mathbf{t}_2)]]] \tag{17}$$

The additional introduced predictive decoder network results in one extra objective in the variational lower bound, and this extra objective can be expressed as:

$$\mathbb{E}_{\mathbf{t} \sim p_\theta(\mathbf{t}|\mathbf{v})}[\log p_\theta(r | \mathbf{t}, \mathbf{v})] = \|r - \hat{r}'\|_2^2 \tag{18}$$

In summary, there are four networks are introduced to approximate multiple conditional probability distributions. These networks include one generative encoder network, one conditional prior encoder network, and two predictive decoder networks. By summing Eq. (15), Eq. (16), and Eq. (18), the total losses of these four networks can be written as Eq. (19), which is a sum of three losses.

$$\begin{aligned} L &= L_1 + L_2 + L_3; \\ L_1 &= \|r - \hat{r}\|_2^2; L_2 = \|r - \hat{r}'\|_2^2; \\ L_3 &= \frac{1}{2} \left(\text{tr}(\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \right. \\ &\left. \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - d + \log \left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \right) \right) \end{aligned} \tag{19}$$

Next, these four networks are connected to train the parameters and extract the features of the surface topography. We name these connected networks as conditional probabilistic autoencoders. Figure 2 shows the flow diagram of the proposed conditional probabilistic autoencoders, where Π is a collection of trainable parameters in the generative encoder network, Π' is a collection of trainable parameters in the conditional prior encoder network, Φ and Φ' refer to collections of trainable parameters in predictive decoder networks.

In the training phase, the process variables \mathbf{v} and MRR r are fed into the generative encoder network to derive $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$; $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ are used to generate the features of the surface topography \mathbf{t}_1 ; both \mathbf{v} and \mathbf{t}_1 are fed into a predictive decoder network to get the predicted MRR \hat{r} . The process variables \mathbf{v} are fed into the conditional prior encoder network to derive

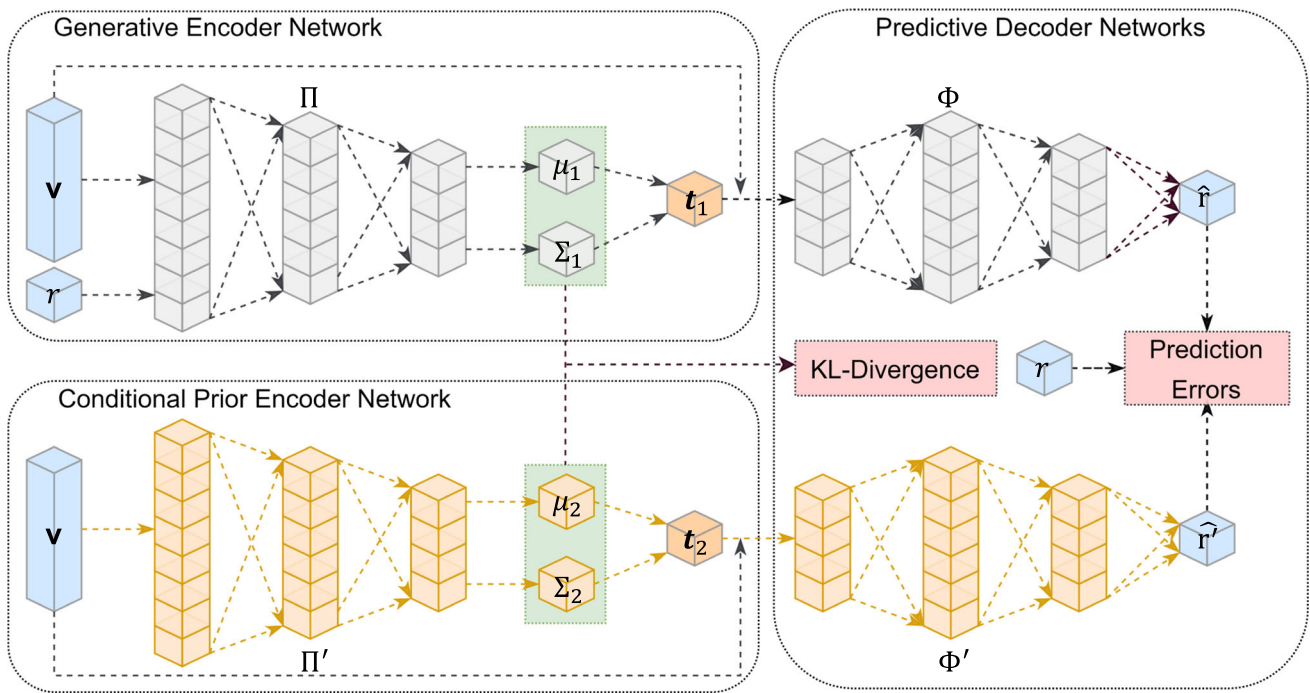


Fig. 2 Flow diagram of the proposed conditional probabilistic autoencoders

μ_2 and Σ_2 ; μ_2 and Σ_2 are used to generate the features of the surface topography t_2 ; both v and t_2 are fed into another predictive decoder network to derive the predicted MRR \hat{r}' ; μ_1 , Σ_1 , μ_2 , and Σ_2 are used to calculate the KL-divergence loss, i.e. L_3 ; r and \hat{r} are used to calculate one prediction loss, i.e. L_1 ; r and \hat{r}' are used to calculate another prediction loss, i.e. L_2 ; Next, all losses L_1 , L_2 , L_3 are back-propagated through these networks to update the trainable parameters, Π , Π' , Φ , and Φ' . In the test phase, process variables are fed into the conditional prior encoder network to extract μ_2 and Σ_2 . μ_2 refers to the deterministic version of the extracted features of the surface topography. The deterministic version of the extracted features helps improve the accuracy of MRR predictions. Table 1 shows the training and test phases of the proposed conditional probabilistic autoencoders.

MRR predictive model

Next, process variables and the extracted features of the surface topography are fed into an ensemble learning-based MRR predictive model to predict the MRR during the CMP process. Ensemble learning usually achieves the best prediction performance by combining multiple base learning algorithms (Polikar, 2006). In this work, we select the best three base regressors out of ten base regressors, the selected base regressors include Random Forests (RF), Gradient Boosting Trees (GBT), and Adaptive Boosting (AB). More details on why these three base regressors were selected

are provided in Sect. **Feature extraction and hyperparameters tuning**. Moreover, the stacking method is implemented to combine three base regressors. These base regressors are briefly introduced in the following sections.

Random forests

The RF refers to an ensemble learning methodology by constructing and combining multiple decision trees (Breiman, 2001; Wu et al., 2019). To develop a decision tree, a random set of variables are selected to split a parent node into two child nodes. The splitting criteria of each parent node can be expressed as the following optimization problem,

$$\min_{j,c} \left[\min_{m_1,m_2} \left(\sum_{x_i \in R_1} (y_i - m_1)^2 + \sum_{x_i \in R_2} (y_i - m_2)^2 \right) \right] \quad (20)$$

where $R_1 = \{x \mid x_j \leq c\}$ and $R_2 = \{x \mid x_j \geq c\}$ refer to two regions after the splitting process is completed; x_j is the j -th splitting variable; c refers to a cutting point; m_1 denotes the mean of the y_i 's that lie into the region R_1 ; and m_2 denotes the mean of the y_i 's that lie into the region R_2 .

The splitting process is replicated until the stopping criteria has been satisfied. A final prediction is made by averaging predictions made by all constructed decision trees.

Table 1 The training and test phases of the proposed conditional probabilistic autoencoders**Methodology:** Conditional Probabilistic Autoencoders to Extract Topography Features**Procedure** Autoencoders-Training

- 1: Construct r and \mathbf{v} , Define Learning rate α
- 2: Randomly initialize Φ , Π , Φ' , Π'
- 3: **while** not done **do**
- 4: Use \mathbf{v} , r to derive μ_1 , Σ_1
- 5: Use \mathbf{v} to derive μ_2 , Σ_2
- 6: Derive \mathbf{t}_1 , \mathbf{t}_2 with μ_1 , Σ_1 , μ_2 , Σ_2
- 7: Predict \hat{r} and \hat{r}' with \mathbf{t}_1 , \mathbf{t}_2 , and \mathbf{v}
- 8: Compute Losses L_1 , L_2 , and L_3
- 9: Update $\Pi \leftarrow \Pi - \alpha \cdot \frac{\partial(L_1+L_3)}{\partial\Pi}$
- 10: Update $\Pi' \leftarrow \Pi' - \alpha \cdot \frac{\partial(L_2+L_3)}{\partial\Pi'}$
- 11: Update $\Phi \leftarrow \Phi - \alpha \cdot \frac{\partial L_1}{\partial\Phi}$
- 12: Update $\Phi' \leftarrow \Phi' - \alpha \cdot \frac{\partial L_2}{\partial\Phi'}$
- 13: **end while**
- 14: **return** parameters Π'

End Procedure**Procedure** Autoencoders-Testing

- 1: Use conditional prior encoder network
- 2: Feed process variables \mathbf{v} to extract μ_2 , Σ_2
- 3: Use μ_2 as the surface topography features

End Procedure**Gradient boosting trees**

The GBT is an ensemble learning method by constructing decision trees sequentially (Friedman, 2001). Higher weights are assigned on data points that are challenging to predict to improve predictive accuracy. The GBT predictor aims at estimating a mapping function $g(\mathbf{x})$ of input \mathbf{x} . The mathematical model is used to approximate the function $g(\mathbf{x})$, which can be written as a sum of multiple local functions g_n ,

$$g(\mathbf{x}) = \sum_{n=1}^N \beta_n g_n(\mathbf{x}; \gamma_n) \quad (21)$$

where N is the total number of local functions; γ_n is a collection of parameters of the local function g_n ; β_n refers to the weighted coefficient. γ_n and β_n can be approximated with the Eq. (22), where L is the loss function.

$$(\gamma_n, \beta_n) = \underset{\gamma, \beta}{\operatorname{argmin}} \mathbb{E}[L(y_i, g_{n-1}(\mathbf{x}_i) + \beta g(\mathbf{x}_i; \gamma))] \quad (22)$$

Next, the greedy-stagewise method (Friedman, 2001) can be used to solve this optimization problem and update the parameters sequentially.

Adaptive boosting

The AB algorithm is similar to the GBT algorithm, which starts from fitting a regressor on the initial dataset and fitting extra regressors on the same dataset with higher weights on data points that are challenging to predict. One primary difference between the AB and GBT algorithms is that the AB algorithm identifies weak learners by high-weight data points, the GBT algorithm identifies weak learners by the gradient. More details of the AB algorithm can be found in Kégl (2013) and Friedman et al. (2000).

Next, the stacking method is employed to combine these three base regressors. A stacking ensemble learning method includes two stages that are training base regressors and training a meta-regressor (Li et al., 2019). Figure 3 shows the two-stage stacking ensemble learning method. In the first stage, process variables and the extracted features of the surface topography are fed into three base regressors to make three individual predictions. In the second stage, these individual predictions are fed into a meta regressor to make a final prediction. In this work, the multi-layer perceptron was employed as a meta regressor.

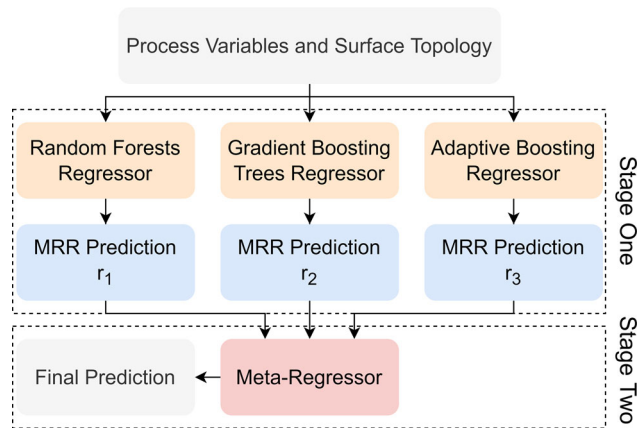


Fig. 3 Two-stage ensemble learning method with stacking

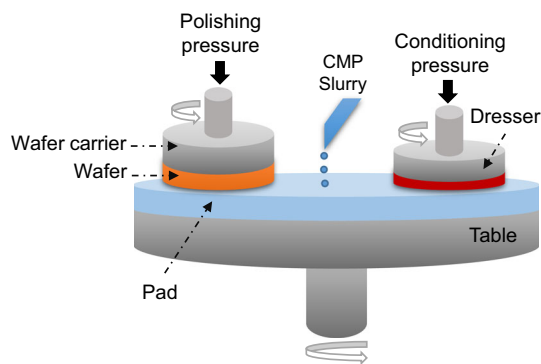


Fig. 4 The schematic diagram of the CMP process

Case study

In this section, the effectiveness of the proposed methodology is demonstrated on a CMP dataset from the PHM data challenge (Li et al., 2019).

Data description

This dataset includes multiple sensor measurements obtained from a CMP process. Figure 4 exhibits a schematic diagram of a typical CMP process. For a typical CMP process, a wafer is captured by a wafer carrier, a polishing pad is attached to the rotating table. In the CMP process, a wafer is pushed toward the planarization pad, and both the rotating table and the wafer carrier are rotated in an identical direction. The abrasive materials are dispensed on the planarization pad via a slurry during the polishing process. A rotating dresser may be engaged in conditioning the polishing pad after the CMP process.

The data includes 19 process variables. These variables, such as chamber pressure, flow rate of slurry, and applied pressure, are real-time collected data. Table 2 lists the symbol and descriptions of these process variables. The real-time

Table 2 Data descriptions

Symbol	Description
v_1	Usage of polish-pad backing film
v_2	Usage of dresser
v_3	Usage of polishing table
v_4	Usage of dresser table
v_5	Chamber pressure
v_6	Pressure applied to the main outer air bag
v_7	Pressure applied to the center air bag
v_8	Pressure applied to the retainer ring
v_9	Pressure applied to the ripple air bag
v_{10}	Usage of polishing membrane
v_{11}	Usage of wafer carrier sheet
v_{12}	Flow rate of slurry type A
v_{13}	Flow rate of slurry type B
v_{14}	Flow rate of slurry type C
v_{15}	Rotating rate of wafer
v_{16}	Rotating rate of stage
v_{17}	Rotating rate of head
v_{18}	Status of dressing water
v_{19}	Pressure applied to the edge air bag

Table 3 Number of wafers in training, validation, and test datasets under two stages

Number of wafers	2016 PHM CMP datasets		
	Training	Validation	Test
Total	1,752	384	378
Stage A	937	212	192
Stage B	815	172	186

data were obtained from wafers under two operating stages (Stage A and Stage B), which are grouped into three datasets, including a training dataset, a validation dataset, and a test dataset. In this work, we remove wafers with a large proportion of missing values to better evaluate the performance of the proposed methodology. Table 3 shows the number of wafers was polished in three datasets under two stages. The proposed method was trained on the training dataset and evaluated on the remaining two datasets.

Feature extraction and hyperparameters tuning

In the previous study, we have demonstrated that five temporal features extracted from the raw data can be used to predict the MRR in the CMP process accurately (Yu et al., 2019). In this case study, we extracted the similar temporal features used in Li et al. (2019) and Yu et al. (2019). The extracted temporal features include mean, median, mode, central moment, and standard deviation; and a total of 95

Table 4 The network structure of the generative encoder network

No. of layers	Description	Output dimension
1	Input layer	batch × 101
2–4	FC & Dropout	batch × 100
5	Mean Generation	batch × 5
6	Variance Generation	batch × 5
7	Features Generation	batch × 5

Table 5 The network structure of the conditional prior encoder network

No. of layers	Description	Output dimension
1	Input layer	batch × 100
2–4	FC & Dropout	batch × 100
5	Mean Generation	batch × 5
6	Variance Generation	batch × 5
7	Features Generation	batch × 5

Table 6 The network structure of the predictive decoder networks

No. of layers	Description	Output dimension
1–3	FC & Dropout	batch × 100
4	FC layer	batch × 1

features were extracted for 19 process variables. Then, the extracted features and the true MRR were fed into the proposed deep probabilistic autoencoder to extract the features of the surface topography. To optimize the performance of the deep probabilistic autoencoder as well as reduce the computational cost, the number of hidden layers in both encoder and decoder networks is set as 3. A dropout layer was added after each hidden layer to avoid the over-fitting problem. The rectified linear unit (ReLU) was used as the activation function in the hidden layers. Because five temporal features were extracted from each process variable, the dimension of the extracted features of the surface topography is also set as 5. Therefore, there is a total of 100 features (95 temporal features and 5 topography features) used for MRR prediction. Tables 4, 5, and 6 show the network structure of the generative encoder network, conditional prior encoder network, and predictive decoder networks. In these tables, batch refers to the batch size, the batch size equals 937 and 815 for wafers manufactured under stage A and stage B, respectively; FC refers to the fully connected layers and Dropout refers to the dropout layers.

Next, we selected the base regressors. It has been demonstrated that combining base regressors of different types can improve the performance of ensemble learning models (Shi et al., 2021). Therefore, we created a base-regressor pool with ten different base regressors, including RF, AB, GBT,

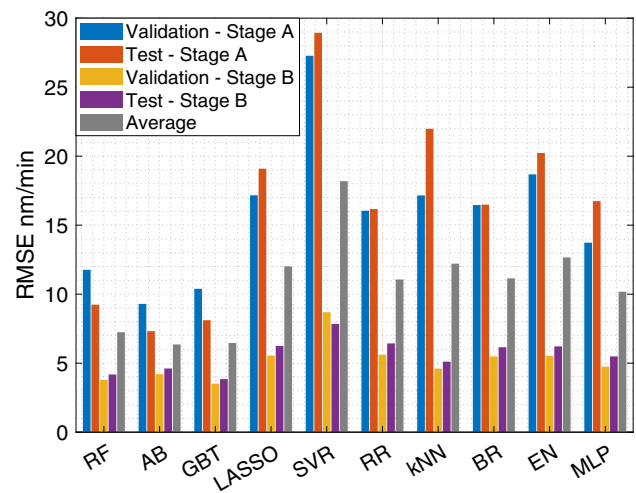


Fig. 5 The RMSE of ten different base regressors for wafers manufactured under two stages on both validation and test dataset

Table 7 The average RMSE of the selected base regressors with respect to different number of estimators

Method	Number of estimators				
	100	200	300	400	500
RF	7.409	7.242	7.255	7.284	7.319
AB	6.356	6.615	6.644	6.399	6.523
GBT	6.617	6.496	6.464	6.470	6.469

LASSO, support vector regression (SVR), ridge regression (RR), k-nearest neighbors (kNN), Bayesian regression (BR), Elastic-Net (EN), and multiple layer perceptron (MLP). In this case study, the best three base regressors were selected to construct the ensemble learning model. Figure 5 shows the RMSE of the ten different base regressors on both validation and test datasets. The results have shown that RF, AB, and GBT are the best three base regressors. Therefore, RF, AB, and GBT were selected as the base regressors in Sect. MRR predictive model.

To optimize the performance of the ensemble learning model, hyperparameter tuning was performed for the selected base regressors. Table 7 shows the average RMSE of the selected base regressors with respect to the different number of estimators. Based on this table, the number of decision trees used in the RF method is set as 200; the number of estimators used in the AB method is set as 100; the number of estimators used in the GBT method is set as 300. Moreover, the meta-regressor is the MLP method which uses 5 hidden layers for simplicity and 100 hidden nodes in each layer in order to be consistent with the number of features used for MRR prediction.

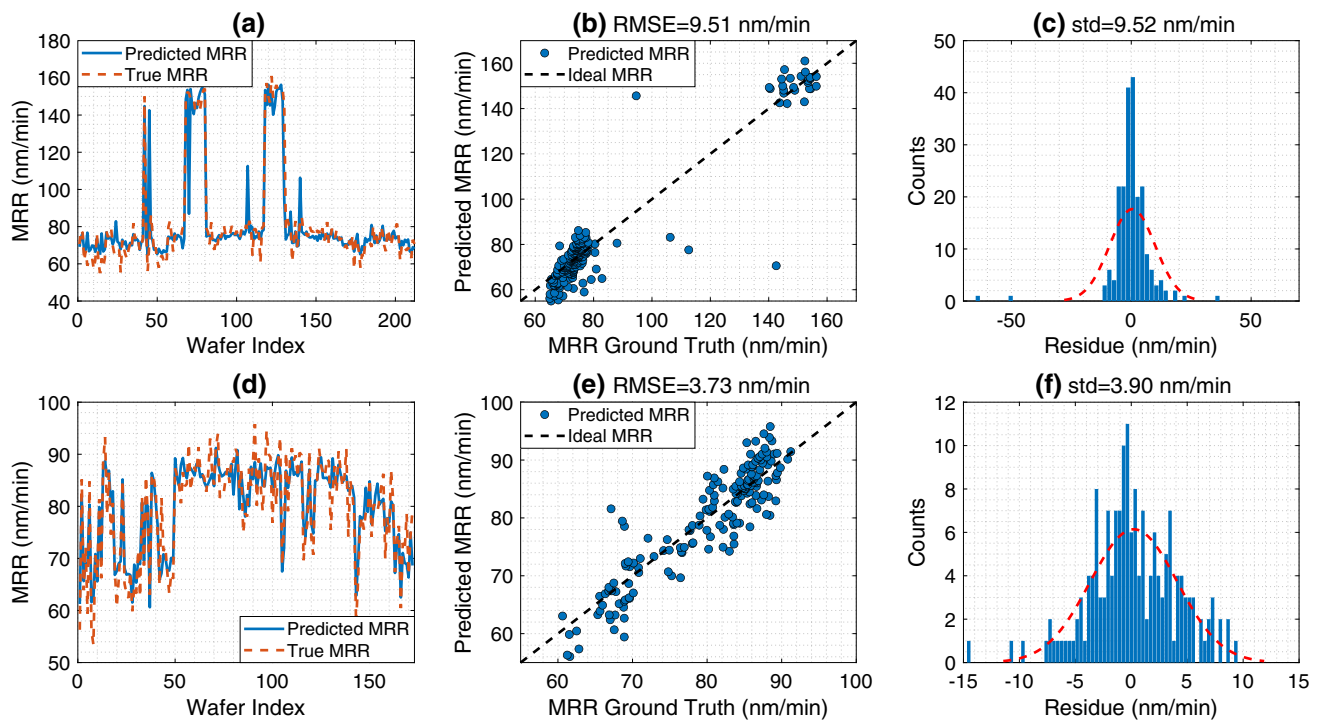


Fig. 6 Prediction performance on validation dataset under two stages. **a–c** show the prediction performance under stage A, **d–f** show the prediction performance under stage B

Results

In this case study, the root mean squared error (RMSE) was used as the performance metric to evaluate the MRR prediction performance. RMSE on validation and test datasets can be calculated using Eq. (23),

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - \hat{r}_i)^2} \quad (23)$$

where r_i and \hat{r}_i are the true and predicted MRR, respectively; and N refers to the total number of wafers.

Figure 6 shows the prediction performance of the proposed methodology on the validation dataset under two polishing stages. Figure 6a–c show the prediction performance under polishing stage A, where Fig. 6a compares the predicted MRR and the true MRR (the ground truth of MRR) in the order of wafer index; Fig. 6b compares the predicted MRR and the true MRR in the order of the MRR; and Fig. 6c presents the histogram and distribution of the prediction difference between the predicted and true MRR. The RMSE of the predicted MRR under polishing stage A is 9.51 nm/min and the standard deviation of the predicted residuals is 9.52 nm/min. Figure 6d–f shows the prediction performance of polishing stage B, where Fig. 6d compares the predicted MRR and the true MRR in the order of wafer index; Fig. 6e compares the predicted MRR and the true MRR in the order

of the MRR; and Fig. 6f shows the histogram and distribution of the prediction difference between the predicted and true MRR. The root mean squared error (RMSE) of the predicted MRR under polishing stage B is 3.73 nm/min and the standard deviation of the predicted residuals is 3.90 nm/min.

Figure 7 shows the prediction performance of the proposed methodology on the test dataset under two polishing stages. Figure 7a–c shows the prediction performance under polishing stage A, where Fig. 7a compares the predicted MRR and the true MRR in the order of wafer index; Fig. 7b compares the predicted MRR and the true MRR in the order of the MRR; and Fig. 7c shows the histogram and distribution of the prediction difference between the predicted and true MRR. The root mean squared error (RMSE) of the predicted MRR under polishing stage A is 7.01 nm/min and the standard deviation of the predicted residuals is 7.72 nm/min. Figure 7d–f shows the prediction performance under polishing stage B, where Fig. 7d compares the predicted MRR and the true MRR in the order of wafer index; Fig. 7e compares the predicted MRR and the true MRR in the order of the MRR; and Fig. 7f shows the histogram and distribution of the prediction difference between the predicted and true MRR. The RMSE of the predicted MRR under polishing stage B is 4.21 nm/min and the standard deviation of the predicted residuals is 4.27 nm/min. Based on these figures, we can observe that wafers polished under Stage A has a higher prediction RMSE and a higher standard deviation in comparison with wafers pol-

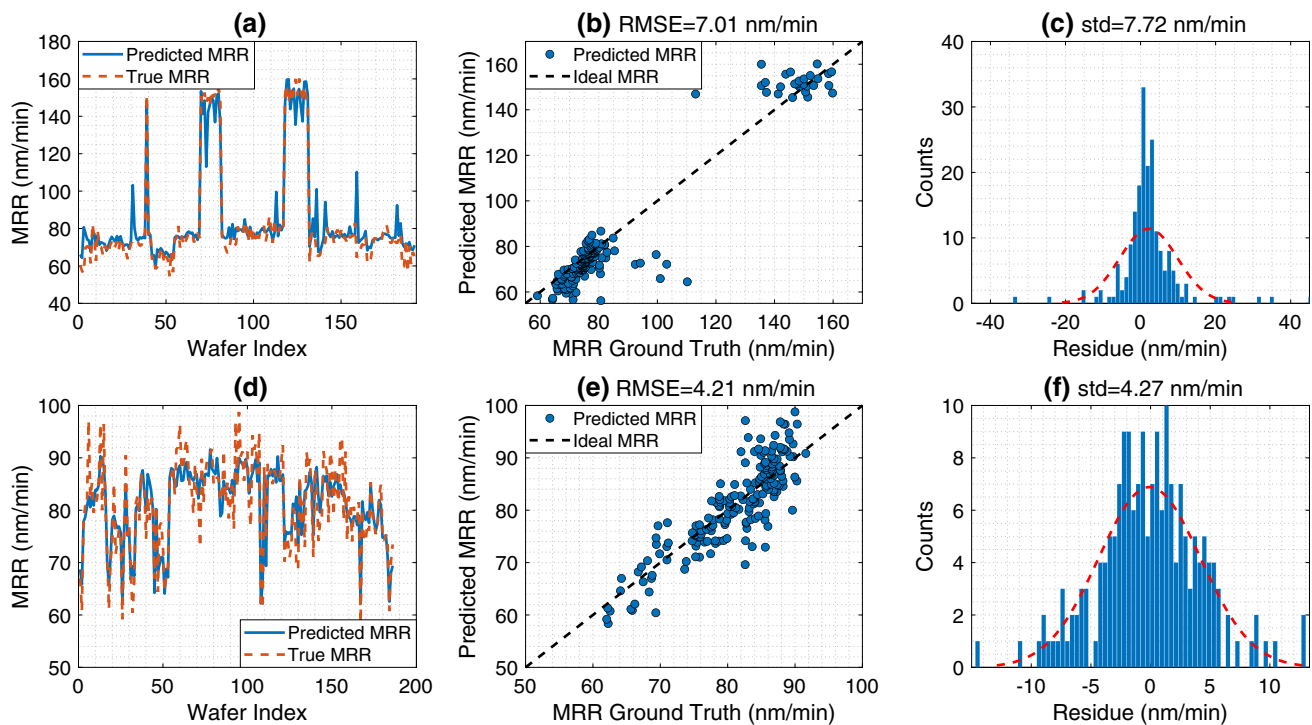


Fig. 7 Prediction performance on test dataset under two stages. **a–c** show the prediction performance under stage A, **d–f** show the prediction performance under stage B

ished under Stage B. One reason that wafers polished under Stage A has a higher prediction RMSE and a higher standard deviation is that the MRR of Stage A is higher than the MRR of Stage B, and the higher MRR brings additional uncertainties for MRR predictions in the CMP process. We can also observe that the prediction residues for both Stage A and Stage B follow normal distributions and has a mean of zero, which means that the proposed method predicts the MRR without underestimations or overestimations. Moreover, operating conditions resulting in lower MRR should be adopted for the CMP process so that wafer-to-wafer thickness variation can be reduced.

Table 8 shows the prediction performance with and without using the extracted surface topography features in terms of RMSE. This table shows that the extracted surface topography features enable a better prediction performance. For example, the RMSE of the MRR predicted without using the extract surface topography features on the test dataset under polishing stage A is 8.25 nm/min. However, the RMSE of the MRR predicted with using the extract surface topography features on the test dataset under polishing stage A is only 7.01 nm/min.

To further demonstrate the effectiveness of the proposed method, the proposed method is also compared with the data-driven methods reported in the literature. Table 9 shows a comparison between the proposed method and other methods reported in the literature in terms of the average RMSE

for both validation and test datasets. The average RMSE refers to the mean of the RMSE of the validation dataset and the RMSE of the test dataset. Based on Table 9, we can conclude that the proposed method outperforms the existing physics-based, data-driven, and physics-informed machine learning models reported in the literature. For example, the average RMSE of the method used in Wang et al. (2017) is 7.60 nm/min. However, the average RMSE of the proposed method is only 6.12 nm/min.

Conclusions and future work

In this paper, a directed graphical model was developed to reveal the relationship among surface topography, process variables, and MRR in the CMP process. Based on the proposed directed graphical model, a deep probabilistic autoencoder was introduced to extract the features of the surface topography. Process variables and the extracted features of surface topography were fed into an ensemble learning-based predictive model to predict the MRR. A CMP dataset was used to demonstrate the effectiveness of the proposed method. The experimental results have shown that the MRR prediction performance can be improved by using the extracted features of the surface topography. The proposed method accurately predicted the MRR in the CMP process with a RMSE of 6.12 nm/min. Moreover, the proposed

Table 8 The prediction performance with and without using the extracted surface topography features in terms of RMSE (nm/min)

	With surface topography		Without surface topography	
	Stage A	Stage B	Stage A	Stage B
Validation	9.51	3.73	10.49	3.55
Test	7.01	4.21	8.25	3.99
Average	6.12		6.57	

Table 9 The prediction performance between the proposed method and other methods reported in the literature

Approach	RMSE (nm/min)
Preston Model (Jia et al., 2018)	29.50
Physics-informed machine learning (Yu et al., 2019)	16.97
Luo and Dornfeld Model (Wang et al., 2017)	7.60
ELM-stacking (Li et al., 2019)	7.24
CART-stacking (Li et al., 2019)	6.97
Proposed Method	6.12

method outperforms existing predictive models reported in the literature in terms of RMSE. In the future, we will consider the dynamic changes in surface topography and their impacts on the MRR predictions. Moreover, different base regressors and ensemble learning methods will also be explored for MRR predictions.

Declarations

Conflict of interest The authors declare no conflict of interest in the research work presented.

References

- Airoldi, E. M. (2007). Getting started in probabilistic graphical models. *PLoS Computational Biology*, 3(12), e252.
- Awano, Y. (2006). Carbon nanotube (cnt) via interconnect technologies: Low temperature cvd growth and chemical mechanical planarization for vertically aligned cnts. In *Proc. 2006 ICPT* (Vol. 10).
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chen, H., Teng, Z., Guo, Z., & Zhao, P. (2020). An integrated target acquisition approach and graphical user interface tool for parallel manipulator assembly. *Journal of Computing and Information Science in Engineering*, 20(2), 021006.
- Deng, J., Zhang, Q., Lu, J., Yan, Q., Pan, J., & Chen, R. (2021). Prediction of the surface roughness and material removal rate in chemical mechanical polishing of single-crystal sic via a back-propagation neural network. *Precision Engineering*, 72, 102–110.
- Evans, C., Paul, E., Dornfeld, D., Lucca, D., Byrne, G., Tricard, M., et al. (2003). Material removal mechanisms in lapping and polishing. *CIRP Annals*, 52(2), 611–633.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.
- Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2), 337–407.
- Greenwood, J. A., & Williamson, J. P. (1966). Contact of nominally flat surfaces. *Proceedings of the royal society of London. Series A: Mathematical and Physical Sciences*, 295(1442), 300–319.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feed-forward networks are universal approximators. *Neural Networks*, 2(5), 359–366.
- Jia, X., Di, Y., Feng, J., Yang, Q., Dai, H., & Lee, J. (2018). Adaptive virtual metrology for semiconductor chemical mechanical planarization process using gmdh-type polynomial neural networks. *Journal of Process Control*, 62, 44–54.
- Johnson, K. L., & Johnson, K. L. (1987). *Contact mechanics*. Cambridge University Press.
- Kégl, B. (2013). The return of adaboost. mh: multi-class hamming trees. arXiv preprint [arXiv:1312.6086](https://arxiv.org/abs/1312.6086).
- Kong, Z., Oztekin, A., Beyca, O. F., Phatak, U., Bukkapatnam, S. T., & 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), 316–327.
- Krishnan, M., Nalaskowski, J. W., & Cook, L. M. (2010). Chemical mechanical planarization: Slurry chemistry, materials, and mechanisms. *Chemical Reviews*, 110(1), 178–204.
- Lee, D., Lee, H., & Jeong, H. (2016). Slurry components in metal chemical mechanical planarization (cmp) process: A review. *International Journal of Precision Engineering and Manufacturing*, 17(12), 1751–1762.
- Lee, H., & 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), 395–403.
- Lee, H., Jeong, H., & Dornfeld, D. (2013). Semi-empirical material removal rate distribution model for sio2 chemical mechanical polishing (cmp) processes. *Precision Engineering*, 37(2), 483–490.
- Lee, K. B., & Kim, C. O. (2020). Recurrent feature-incorporated convolutional neural network for virtual metrology of the chemical mechanical planarization process. *Journal of Intelligent Manufacturing*, 31(1), 73–86.
- Leon, J. I., Vazquez, S., & Franquelo, L. G. (2017). Multilevel converters: Control and modulation techniques for their operation and industrial applications. *Proceedings of the IEEE*, 105(11), 2066–2081.
- Li, Z., Wu, D., & Yu, T. (2019). Prediction of material removal rate for chemical mechanical planarization using decision tree-based

- ensemble learning. *Journal of Manufacturing Science and Engineering*, 141(3), 031003.
- Luo, J., & Dornfeld, D. A. (2001). Material removal mechanism in chemical mechanical polishing: theory and modeling. *IEEE Transactions on Semiconductor Manufacturing*, 14(2), 112–133.
- Luo, Q., Ramarajan, S., & Babu, S. (1998). Modification of the Preston equation for the chemical-mechanical polishing of copper. *Thin Solid Films*, 335(1–2), 160–167.
- Nguyen, N., Zhong, Z., & Tian, Y. (2015). An analytical investigation of pad wear caused by the conditioner in fixed abrasive chemical-mechanical polishing. *The International Journal of Advanced Manufacturing Technology*, 77(5), 897–905.
- Oh, S., & Seok, J. (2009). An integrated material removal model for silicon dioxide layers in chemical mechanical polishing processes. *Wear*, 266(7–8), 839–849.
- Pandey, G., & Dukkipati, A. (2017). Variational methods for conditional multimodal deep learning. In *2017 international joint conference on neural networks (IJCNN)* (pp. 308–315). IEEE.
- Park, B., Lee, H., Park, K., Kim, H., & Jeong, H. (2008). Pad roughness variation and its effect on material removal profile in ceria-based cmp slurry. *Journal of Materials Processing Technology*, 203(1–3), 287–292.
- Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3), 21–45.
- Sheu, D. D., Chen, C.-H., & Yu, P.-Y. (2012). Invention principles and contradiction matrix for semiconductor manufacturing industry: chemical mechanical polishing. *Journal of Intelligent Manufacturing*, 23(5), 1637–1648.
- Shi, J., Yu, T., Goebel, K., & Wu, D. (2021). Remaining useful life prediction of bearings using ensemble learning: The impact of diversity in base learners and features. *Journal of Computing and Information Science in Engineering*, 21(2).
- Steigerwald, J. M., Murarka, S. P., & Gutmann, R. J. (1997). *Chemical mechanical planarization of microelectronic materials*. Wiley.
- Wang, P., Gao, R. X., & Yan, R. (2017). A deep learning-based approach to material removal rate prediction in polishing. *CIRP Annals*, 66(1), 429–432.
- Wei, Y., Wu, D., & Terpenney, J. (2021). Learning the health index of complex systems using dynamic conditional variational autoencoders. *Reliability Engineering & System Safety*, 216, 108004.
- Wu, D., Wei, Y., & Terpenney, J. (2019). Predictive modelling of surface roughness in fused deposition modelling using data fusion. *International Journal of Production Research*, 57(12), 3992–4006.
- Xia, L., Zheng, P., Huang, X., & Liu, C. (2021). A novel hypergraph convolution network-based approach for predicting the material removal rate in chemical mechanical planarization. *Journal of Intelligent Manufacturing*. <https://doi.org/10.1007/s10845-021-01784-1>
- Yin, S., Rodriguez-Andina, J. J., & Jiang, Y. (2019). Real-time monitoring and control of industrial cyberphysical systems: With integrated plant-wide monitoring and control framework. *IEEE Industrial Electronics Magazine*, 13(4), 38–47.
- Yu, T., Asplund, D. T., Bastawros, A. F., & Chandra, A. (2016). Performance and modeling of paired polishing process. *International Journal of Machine Tools and Manufacture*, 109, 49–57.
- Yu, T., Li, Z., & Wu, D. (2019). Predictive modeling of material removal rate in chemical mechanical planarization with physics-informed machine learning. *Wear*, 426, 1430–1438.
- Zantye, P. B., Kumar, A., & Sikder, A. (2004). Chemical mechanical planarization for microelectronics applications. *Materials Science and Engineering: R: Reports*, 45(3–6), 89–220.
- Zhao, T., Zhao, R., & Eskenazi, M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. arXiv preprint [arXiv:1703.10960](https://arxiv.org/abs/1703.10960).
- Zhao, Y., & Chang, L. (2002). A micro-contact and wear model for chemical-mechanical polishing of silicon wafers. *Wear*, 252(3–4), 220–226.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.