MTGPVM: Adaptive Active Learning Assisted Virtual Metrology Modeling in Chemical Vapor Deposition Systems

Shanling Ji
Southeast University
Min Dai
Southeast University
Haiying Wen
Southeast University
Hui Zhang
Southeast University
Zhisheng Zhang
Southeast University
Zhijie Xia
Southeast University
Jianxiong Zhu (danverzhu@gmail.com)
Southeast University https://orcid.org/0000-0002-9172-5255

Research Article

Keywords: Active learning, Adaptive learning, Multitask Gaussian process, PECVD, Virtual metrology

Posted Date: April 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1550928/v1

MTGPVM: Adaptive Active Learning Assisted Virtual Metrology Modeling in Chemical Vapor Deposition Systems

Shanling Ji [1], Min Dai [1], Haiying Wen [1], Hui Zhang [1], Zhisheng Zhang [1]*, Zhijie Xia [1]*, Jianxiong Zhu [1,2]*

[1]School of mechanical engineering, Southeast University, Nanjing, 211189, P. R. China.

[2]State Key Laboratory of Transducer Technology, Chinese Academy Sciences, Shanghai, 200050, P. R. China

*Corresponding author. Email: oldbc@seu.edu.cn; zxia@seu.edu.cn; mezhujx@seu.edu.cn.

Abstract: Virtual metrology estimates the quality using processing variables collected from the multi-sensor system and enhances the detection efficiency compared with manual real metrology. However, the multitask learning problem and few labeled sets from available real metrology are challenges for virtual metrology modeling. Accordingly, this paper presents the improved method to combine multi-task Gaussian process and adaptive active learning for the application of virtual metrology in intelligent manufacturing. Initially, a multitask Gaussian processes-based virtual metrology (MTGPVM) model is built by combining the coregionalization matrix and two basic kernel functions and sharing latent correlation characteristics among tasks. Furthermore, three query functions utilizing uncertain sampling and diversity sampling are proposed for the active learning of multitask virtual metrology. Additionally, the adaptive active learning algorithm is proposed for virtual metrology model training with the few labeled data of real metrology. Finally, the case study on the proposed methods is carried out based on the coating processing of solar silicon wafers with the technology of plasma-enhanced chemical vapor deposition (PECVD). The experimental results indicate that the MTGPVM model is based on the combination of outperformed Martern kernel and linear kernel other multitask regression models, the proposed adaptive active learning algorithm prompted the learning accuracy of MTGPVM model with a small amount of labeled training data.

Keywords: Active learning; Adaptive learning; Multitask Gaussian process; PECVD; Virtual metrology.

1. Introduction

Virtual metrology (VM) is a method to augment existing metrology and can be used in the control of the process to improve processing accuracy and speed. The VM model is built with previous real metrology data and the processing quality can be predicted by VM model with processing variables and processing states. Especially in semiconductor manufacturing, the metrology results have been utilized in the run-to-run controller to improve the processing recipes of the next product batch [1]. In actual engineering, the physical metrology is difficult to be implemented for all processing wafers limited to the actual operation conditions, such as additional cost, human resources, and long cycle time. Additionally, an imprecise VM model may misguide the feedback control module. Therefore, establishing a VM model with low errors and high reliability is essential for processing control in intelligent manufacturing.

Prior VM modeling research are based on learning methods including machine learning methods and deep learning methods. In the terms of deep learning, neural networks with different structures are utilized in VM modelings, such as convolutional neural networks [2, 3] and recurrent neural networks [4]. In some cases, the composition of processing variables collected by sensors may be easily expressed such that using deep learning may introduce unnecessary training costs or even overfitting. Machine learning models for VM include support vector regression (SVR) [5], linear regression, tree bagging [6], k-Nearest Neighbor (k-NN) [7], and Gaussian process regression (GPR) [8]. GPR is a nonparametric method based on the Bayesian modeling technique. Compared with other machine learning models, GPR for VM modeling is widely researched for the reason that the engineering manufacturing process can be approximated as a complex combination of Gaussian processes.

Multitask learning can learn multiple related tasks by sharing latent information and can enhance generalization performance. There are usually multiple quality factors to be monitored in the manufacturing process. Thus, the VM model based on multitask learning can be applied in multitask manufacturing scenarios. Multitask Gaussian process is a multitask learning model that uses two covariance functions to correlate information between different tasks and different inputs and indicate multi-output sets with the vector-valued function [9, 10]. Although work [11] has compared the VM predictive accuracy of Multitask Gaussian processes with

other machine learning methods, the performance of using different kernel alternatives is ignored.

The high cost of labeling equipment and low efficiency of manual labeling often appear in real metrology, which results in VM learning with small amounts of manually annotated data. Active learning is a method that can address the learning problem with small amounts of labeled data and further large amounts of unlabeled data. Different from passive learning training model based on abundant annotated samples, active learning can train the model with fewer annotated samples and a larger amount of unlabeled samples. With uncertainty sampling, active learning has been proposed to improve prediction accuracy by iteratively updating the VM model in work [12], although the VM modeling of this work is based on neural networks rather than Gaussian processes.

From the literature we reviewed, there is currently no systematic VM framework using multitask Gaussian processes model and active learning for connecting discrete production recipes and quality factors. To solve the multitask and few label data of real metrology, we proposed a multitask Gaussian processes enabled VM modeling method and corresponding adaptive active learning algorithm in this study. The diagram of the intelligent manufacturing system based on the proposed VM modeling method is illustrated in Fig. 1, which contains four parts:

(a) The physical manufacturing process to process products and real metrology to measure quality manually.
(b) Data extraction and fusion for the processing data of physical manufacturing process and quality data of real metrology. The processing data without quality measurement is regarded as an unlabeled set.
(c) The proposed VM modeling methods collect the labeled data and train it based on multitask Gaussian process model and the improved adaptive active learning method.
(d) The VM estimates quality factors based on the proposed learning model and then the quality controller (always a run-to-run controller in semiconductor manufacturing) analyzes the suggestions for the next process.

Therein, a general VM function is proposed to relate the processing variables and multivariable predictive quality. Different combinations of kernel functions are compared using the dataset from real manufacturing processes. After that, the proposed adaptive active learning strategy is implemented for cost-effective training using real metrology data. The rest of the paper is organized as follows. Section II introduces the background knowledge of the basic GPR and Multitask Gaussian process as well as the related work about active learning. The proposed adaptive active learning method with multitask Gaussian processes for VM modeling is detailed in section III. Section IV investigates the performance of the proposed method through a case study on an engineering dataset. Finally, conclusions are presented in section V.

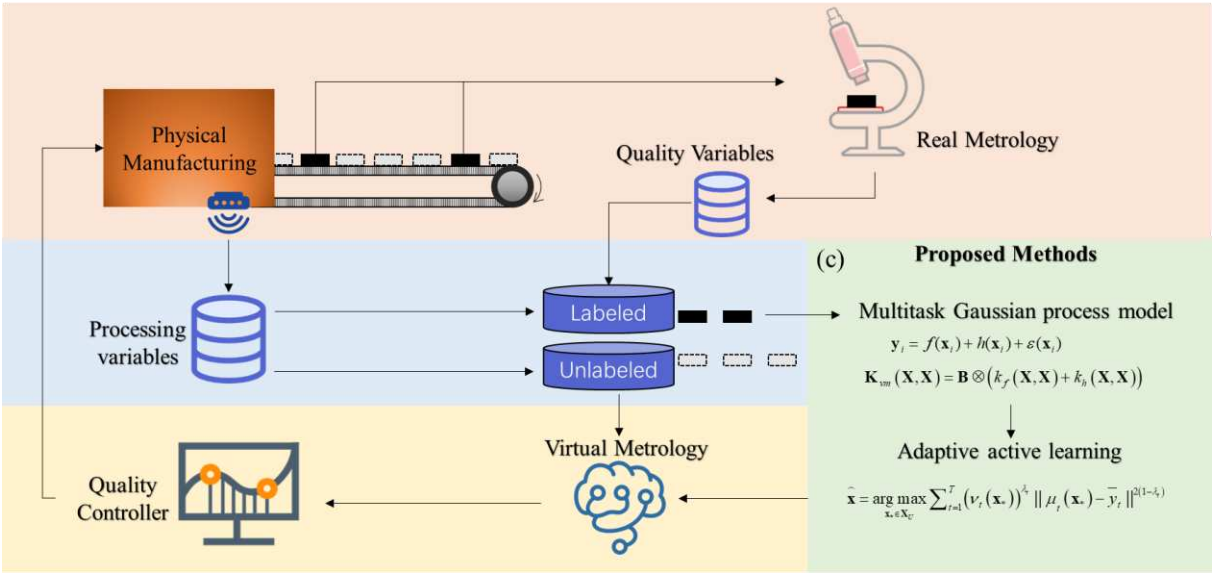

Fig. 1 The diagram of virtual metrology-based intelligent semiconductor manufacturing system. (a) The physical manufacturing process and real metrology. (b) Data extraction and fusion. (c) Proposed VM modeling method based on multitask Gaussian process and improved adaptive active learning. (d) Virtual metrology-based quality feedback controller (run-to-run controller).

2. Background and Related Work

In this section, the background knowledge and related work about the GPR with single output and multi-task Gaussian processes as well as active learning are introduced.

2.1. The GPR with a single output

A Gaussian process is a set of random variables from which any finite number of random variables have a joint Gaussian distribution [13]. Let $S=(\mathbf{X},\mathbf{Y})=\{(\mathbf{x}_i, y_i)\,|\,i=1,...,N\}$ denotes the N pairs of training data points, where $\mathbf{x}$ is the directly observed data and y the corresponding response. A zero-mean Gaussian process can be defined as $f(\mathbf{X}) \sim N(0, k(\mathbf{X},\mathbf{X}'))$, where $k(\mathbf{X},\mathbf{X})$ is the covariance function (also known as the kernel function). Considering the independent white noise, the relation between observed input $\mathbf{x}_i$ and output y_i is given by:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim N(0, \sigma_i^2\mathbf{I})$ is white noise and independent of $f(\mathbf{x}_i)$.

To predict $f_*(\mathbf{X}_*) \sim N(0, k(\mathbf{X}_*,\mathbf{X}_*))$ the positions of new observed points $\mathbf{X}_*$, the joint Gaussian distribution of the training instances and test instances is used and written as:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I} & k(\mathbf{X},\mathbf{X}_*) \\ k(\mathbf{X}_*,\mathbf{X}) & k(\mathbf{X}_*,\mathbf{X}_*) \end{bmatrix}\right) \tag{2}$$

After that, the posterior distribution f_* is given by:

$$p(f_* \,|\, \mathbf{X}_*, y) \sim N(m(\mathbf{X}_*), cov(\mathbf{X}_*)) \tag{3}$$

$$m(\mathbf{X}_*) = k(\mathbf{X}_*,\mathbf{X})\left[k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\right]^{-1} y \tag{4}$$

$$cov(\mathbf{X}_*) = k(\mathbf{X}_*,\mathbf{X}_*) - k(\mathbf{X}_*,\mathbf{X})\left[k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\right]^{-1} k(\mathbf{X},\mathbf{X}_*) \tag{5}$$

The learning process of GPR with training datasets is to optimize hyperparameters of kernel functions. The selection of hyperparameters $\boldsymbol{\theta}$ in the covariance function can be obtained by minimizing the negative logarithmic likelihood function $p(y \,|\, \mathbf{X}, \boldsymbol{\theta})$ [14].

2.2. Multitask Gaussian processes

Compared with the basic Gaussian process model with a single output, the multi-task Gaussian processes can correlate between tasks with extra covariance functions. Let $k(\mathbf{x},\mathbf{x}')$ and $k_t(t,t')$ represent the covariance of input data and covariance between multiple outputs, which

t means two different tasks (assume T tasks in total). The covariance function $K(\mathbf{x},\mathbf{x}') = k_t(t,t') \otimes k(\mathbf{x},\mathbf{x}')$ can recognize the correlations between multiple tasks, where $\otimes$ is the Kronecker product. In terms of the linear model of coregionalization (LMC), the covariance matrix can be expressed as a sum of latent separable kernels:

$$\mathbf{K}(\mathbf{x},\mathbf{x}') = \sum_{l=1}^{L} \mathbf{B}_l \otimes k_l(\mathbf{x},\mathbf{x}') \tag{6}$$

where $\mathbf{B}_l$ means the coregionalization matrix of latent number l. The intrinsic coregionalization model (ICM) is a simplified version of the LMC $L=1$ [10].

The hyperparameters of the multitask Gaussian process model can also be optimized by minimizing the negative logarithmic likelihood function, and the prediction using test data can be made based on conditional probability [15].

2.3. Active learning

Active learning can actively select unlabeled samples with abundant information and hand them to experts for annotation, and then put them into a training set for training, to obtain higher accuracy in the case of a small training set, which can effectively reduce the cost of constructing high-performance VM model. The definition of query function or the selection method for the next sample sequentially is the key to active learning to obtain better results and cope with a lack of data. Uncertain sampling is among the most popular approaches and queries the informative instances with the maximal uncertainty. In terms of VM regression work, the uncertainty can be estimated by the variance of outputs based on neural networks [12] and Gaussian processes [16]. However, it is not conducive to reliable GPR learning to only consider the prediction variance as a single information measure because of the exiting cases where the prediction variance is smaller while the prediction error is larger. Additionally, diversity sampling is another approach to selecting informative instances. Correlatively, L2 distance has been utilized to calculate the diversity in both the input and output spaces for multi-task learning [17]. Considering that the active learning combining diversity sampling and uncertain sampling has not been applied to multitask learning in the above research, especially in the regression based on multitask Gaussian processes, we are motivated to propose and detail our algorithms in the subsequent section.

3. Proposed Approach

The improved active learning-based multitask Gaussian processes method is proposed for VM modeling in this section. Firstly, a general VM function for quality prediction is established based on the coregionalization model. After that, the hyperparameters optimization method is introduced. Afterward, the information measure criteria for active learning based on the proposed VM model are introduced. Finally, adaptive active learning-based training algorithm is improved to solve the high cost and low efficiency of real metrology.

3.1. VM modeling for quality prediction

First, the general VM model based on multi-task Gaussian processes (MTGPVM) is proposed in this study. There is a coupling between the prediction of quality variables and the latent characteristics of processing variables in industrial engineering. The general VM function to simultaneously connect multiple quality factors $\mathbf{y}_i$ with processing variables $\mathbf{x}_i$ from multi-sensor information can be expressed as:

$$\mathbf{y}_i = f(\mathbf{x}_i) + h(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i) \tag{7}$$

where $f(\mathbf{x}_i)$ is zero-mean Gaussian process to extract the covariance of individual characteristic, $h(\mathbf{x}_i)$ represents the basis function to extract deviation information due to different mean values of quality factors. $\varepsilon(\mathbf{x}_i)$ is the white noise with the covariance function of $\sigma_i^2\mathbf{I}$.

The coregionalization model (6) is used to specify the $f(\mathbf{x}_i)$ and $h(\mathbf{x}_i)$ for multi-out covariance functions:

$$\operatorname{cov}\left(f(\mathbf{x}_i), f(\mathbf{x}_j)\right) = \sum_{l_1=1}^{L_1} \mathbf{B}_{l_1} \otimes k_{l_1}\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{8}$$

$$\operatorname{cov}\left(h(\mathbf{x}_i), h(\mathbf{x}_j)\right) = \sum_{l_2=1}^{L_2} \mathbf{B}_{l_2} \otimes k_{l_2}\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{9}$$

where $\mathbf{B}_{l_1}$ and $\mathbf{B}_{l_2}$ are coregionalization matrixes of two covariance functions. $k_{l_1}\left(\mathbf{x}_i,\mathbf{x}_j\right)$ and $k_{l_2}\left(\mathbf{x}_i,\mathbf{x}_j\right)$ are two basic kernel functions. For simply expression with ICM, we assume $L_1 = L_2 = 1$. Then, the covariance functions of $f(\mathbf{x}_i)$ and $h(\mathbf{x}_i)$ can be expressed as:

$$\mathbf{K}_f\left(\mathbf{X},\mathbf{X}\right)+\mathbf{K}_h\left(\mathbf{X},\mathbf{X}\right)=\mathbf{B}_{vm}\otimes\left(k_f\left(\mathbf{X},\mathbf{X}\right)+k_h\left(\mathbf{X},\mathbf{X}\right)\right) \tag{10}$$

where $\mathbf{B}_{vm}$ is a $T\times T$ coregionalization matrix correlating the outputs for different quality factors and sharing latent information from two kernel matrixes. $k_f\left(\mathbf{X},\mathbf{X}\right)$ and $k_h\left(\mathbf{X},\mathbf{X}\right)$ are two basic kernel matrixes. The $k_f\left(\mathbf{X},\mathbf{X}\right)$ measure is correlated by the distance of two inputs. The $k_h\left(\mathbf{X},\mathbf{X}\right)$ can represent offset information in a complex system. Thus, the covariance matrix of MTGPVM is:

$$\mathbf{K}_{vm}\left(\mathbf{X},\mathbf{X}\right)=\mathbf{B}_{vm}\otimes\left(k_f\left(\mathbf{X},\mathbf{X}\right)+k_h\left(\mathbf{X},\mathbf{X}\right)\right) \tag{11}$$

Let $\boldsymbol{\theta}$ indicate hyperparameters of the covariance matrix $\mathrm{K_{VM}}(\mathbf{X},\mathbf{X})$. Then the prediction distribution based on the general MTGPVM model (7) is:

$$p\left(\mathbf{y}_*\,|\,\mathbf{x}_*,\mathbf{x},\mathbf{y},\boldsymbol{\theta}\right)\square\ N\left(\boldsymbol{\mu}\left(\mathbf{x}_*\right),\mathbf{v}\left(\mathbf{x}_*\right)\right) \tag{12}$$

$$\boldsymbol{\mu}\left(\mathbf{x}_*\right)=\mathbf{K}_{vm}\left(\mathbf{x}_*,\mathbf{x}\right)\left[\mathbf{K}_{vm}\left(\mathbf{x},\mathbf{x}\right)+\sigma^2\mathbf{I}\right]^{-1}\mathbf{y} \tag{13}$$

$$\mathbf{v}\left(\mathbf{x}_*\right)=\mathbf{K}_{vm}\left(\mathbf{x}_*,\mathbf{x}_*\right)-\mathbf{K}_{vm}\left(\mathbf{x}_*,\mathbf{x}\right)\left[\mathbf{K}_{vm}\left(\mathbf{x},\mathbf{x}\right)+\sigma^2\mathbf{I}\right]^{-1}\mathbf{K}_{vm}\left(\mathbf{x},\mathbf{x}_*\right) \tag{14}$$

The criterion for prediction with good performance is that the mean vector function $\boldsymbol{\mu}\left(\mathbf{x}_*\right)$ is closer to the real metrology data and the variance vector function $\mathbf{v}\left(\mathbf{x}_*\right)$ has a smaller uncertain bound.

3.2. Hyperparameters optimization

The hyperparameters are initialized to random numbers at the beginning of training and then a multivariate optimization algorithm is used for an iterative learning solution to search for the optimal choice of the hyperparameters. The variational Gaussian processes approximation is utilized by minimizing the KL divergence between the variational distribution and the posterior

distribution, which is equivalent to maximizing the evidence lower bound (ELBO) [18]. The specific procedures based on the natural gradient descent (NGD) in GPflow [19] are implemented for the hyperparameters optimization in this study.

Several common basic kernel functions are considered to demonstrate the relationship among the real metrology data. The squared exponential (SE) covariance function k_{se} is a popular choice:

$$k_{se}(r) = \theta_\alpha^2 \exp\left(-\frac{1}{\theta_l^2} \| r \|^2 \right) \tag{15}$$

where $r = |x - x'|$. θ_α and θ_l are hyperparameters. θ_l determines the scale characteristics of the input. However, the SE kernel cannot describe the tendency of non-stationary changes in random processes because of the complex characters of many physical phenomena. The more common alternative is the Matern kernel function and the choice of smoothness hyperparameter $\nu = 3/2$ is denoted as k_{Ma3}:

$$k_{Ma3}(r) = \left(1 + \frac{\sqrt{3}r}{\theta_m} \right) \exp\left(-\frac{\sqrt{3}r}{\theta_m} \right) \tag{16}$$

where θ_m is the scale hyperparameter.

In addition, bias kernel k_{bias} and linear kernel k_{linear} are always used to further fit bias or linear variation in the real scenario:

$$k_{bias}(x, x') = c_b^2 \tag{17}$$

$$k_{linear}(x, x') = c_l{}^2 x x' \tag{18}$$

where c_b and c_l denote their corresponding hyperparameters.

3.3. Information measure criterion

As mentioned before, unlabeled sets are defined by U the need to be labeled to get a more accurate prediction. Let L mean the labeled set. As labeling data using real metrology is time-consuming, the strategy of actively choosing the most informative predictive data into labeled data is essential. In this study, active learning is exploited for MTGPVM with uncertain sampling and diversity sampling.

First, the adaptive uncertain sampling (AUNs) criterion for the most informative instances with the maximum prediction variance is proposed as:

$$\mathbf{x} = \arg\max_{\mathbf{x}_* \in \mathbf{X}_U} \sum_{t=1}^{T} \left(\nu_t \left(\mathbf{x}_* \right) \right)^{\lambda_t} \tag{19}$$

where λ_t is within the range of [0,1] and means the weight coefficient for task *t*. (19) becomes a query function of the non-adaptive uncertain sampling (UNs) when the weight coefficient λ_t is a constant.

Afterward, the adaptive diversity sampling (ADIs) method is designed to sample the maximum diversity information of regression outputs between the labeled set and the unlabeled set. Let $\overline{y_t}$ denote the reference value of task *t* from the labeled set, then the diversity criterion is proposed as:

$$\mathbf{x} = \arg\max_{\mathbf{x}_* \in \mathbf{X}_U} \sum_{t=1}^{T} \| \mu_t \left(\mathbf{x}_* \right) - \overline{y_t} \|^{2\lambda_t} \tag{20}$$

Similarly, (20) is the non-adaptive diversity sampling (DIs) when weight coefficients are fixed.

Finally, the adaptive active learning method of the combined sampling (ACOs) is expressed as:

$$\mathbf{x} = \arg\max_{\mathbf{x}_* \in \mathbf{X}_U} \sum_{t=1}^{T} \left(\nu_t \left(\mathbf{x}_* \right) \right)^{\lambda_t} \| \mu_t \left(\mathbf{x}_* \right) - \overline{y}_t \|^{2(1-\lambda_t)} \tag{21}$$

Let $\lambda_t = 0.5$ indicates the equal-weight of two sampling methods for the combination sampling (COs). When $0 < \lambda_t < 0.5$, the weight of diversity information is enhanced. Conversely, the weight of uncertain information is stressed when $0.5 < \lambda_t < 1$.

3.4. Adaptive active learning

The mean-absolute-percentage error (MAPE) is used to evaluate the performance of quality prediction in the task *t*:

$$e_t = \frac{\sum_{i=1}^{n} \left| \left(y_{t,i} - y_{t,i} \right) / y_{t,i} \right|}{n} \tag{22}$$

where $\hat{y}_{t,i}$ is the i-th predictive value of t-th quality factor, $y_{t,i}$ is the real metrology value of t-th quality factor, n is the sample size and t is the index of quality factor. The range of MAPE is [0,1] and the value is 1 when the calculated value is greater than or equal to 1.

Assume the unlabeled set is generated from the divided historical data pool or the data stream from the recent real metrology, the purpose is to actively train MTGPVM model using the proposed sampling criterions. Let κ indicates a number of query unlabeled samples. Initially, the labeled set is divided into the training set L_κ and a validation set L_v. The unlabeled set U_κ mainly includes the processing variables. Before the iteration, the initial training set is utilized to train MTGPVM model. For the κ-th iteration, the prediction error $e_{t,\kappa}$ for exiting the training set using (22) is calculated at the beginning and used to adaptively modulate the sampling criterion by updating the weight coefficient of each task:

$$\lambda_{t,\kappa} = \left(1-\alpha e_{t,\kappa}\right)\lambda_{t,\kappa-1} + \alpha e_{t,\kappa} \tag{23}$$

where $\lambda_{t,\kappa}$ means the weight coefficients of sampling criterion in the k-th iteration. α is a learning rate. The unlabeled samples U_s are selected by the query function and assigned with the quality variables from the real metrology. Afterward, the selected samples U_s are added to the training set, which results in the updating of the labeled training set and the unlabeled set. Finally, the MTGPVM model is retrained and hyperparameters of covariance functions are further optimized until the end of the iteration. The termination condition may be the convergence of the model trained with the updated sample or the unlabeled dataset has been sampled completely. The multitask Gaussian processes with adaptive active learning for virtual metrology based on the variational inference and the natural gradient descent are summarized in Algorithm 1.

Algorithm 1 Adaptive active learning algorithm for MTGPVM

Input: Labeled training set L_0, labeled validation set L_v, unlabeled set U_0, the initial weight coefficient $\lambda_{t,0}$.

Output: Active learned multitask Gaussian processes model for virtual metrology.

1. Initialize the MTGPVM model (7) on the labeled set L_0 using the NGD.
2. **While** not termination condition **do**
3. **for** t =1 **to** T **do**
4. Calculate $e_{t,k}$ using exiting labeled training set L_κ with (22).
5. Update weight coefficient $\lambda_{t,\kappa}$ using (23).
6. **end for**
7. Select the unlabeled set of processing variables U_s using (19), (20), or (21).
8. Assign quality variables to the selected unlabeled set.
9. Add the selected samples to the training set: $L_{\kappa+1} \leftarrow L_\kappa \bigcup U_s$.
10. Update the unlabeled set: $U_{\kappa+1} \leftarrow U_\kappa - U_s$.
11. Record MAPE of the trained model using the validation set L_v.
12. Retrain the MTGPVM model with the updated labeled set $L_{\kappa+1}$.
13. **end while**

4. Case Study

4.1. Experimental setup and data preprocessing

Fig. 2 shows the experimental setup of the case study. The experiment data is generated from the processing process of silicon solar-cell antireflection coating. The coating method is plasma-enhanced chemical vapor deposition (PECVD), which deposits silicon nitride (SiNx) films on the solar cell by under various conditions. For the solar cells processed in the same batch and same reaction chamber, the SiNx thickness in different positions can be measured

using the ellipsometer. The surface of the silicon wafer grows from 65nm-80nm after the total coating process in our collected dataset. The thickness of sampled silicon wafers is manually measured for the fixed positions of the processing boat and fix locations of the silicon wafer. The deposition rate for fixed operation conditions has a theoretical value, which makes the surface thickness grow according to the linear accumulation of deposition time. However, because of the gravity and imbalanced distribution of the reaction environment, there are irregular deposition results at different positions in the chamber. Fig. 2(a) is the digital-twin visualized PECVD manufacturing process for real-time quality monitoring. The processing status and real metrology results of sampling points are visualized by the digital model. The virtual metrology fills up the data of unmeasured samples. Fig. 2(b) demonstrated the real manufacturing system of silicon wafers using PECVD technology. Fig. 2(c) shows the diagram of the silicon solar-cell coating with the PECVD processing, the processing variables and quality variables for the VM modeling are represented.

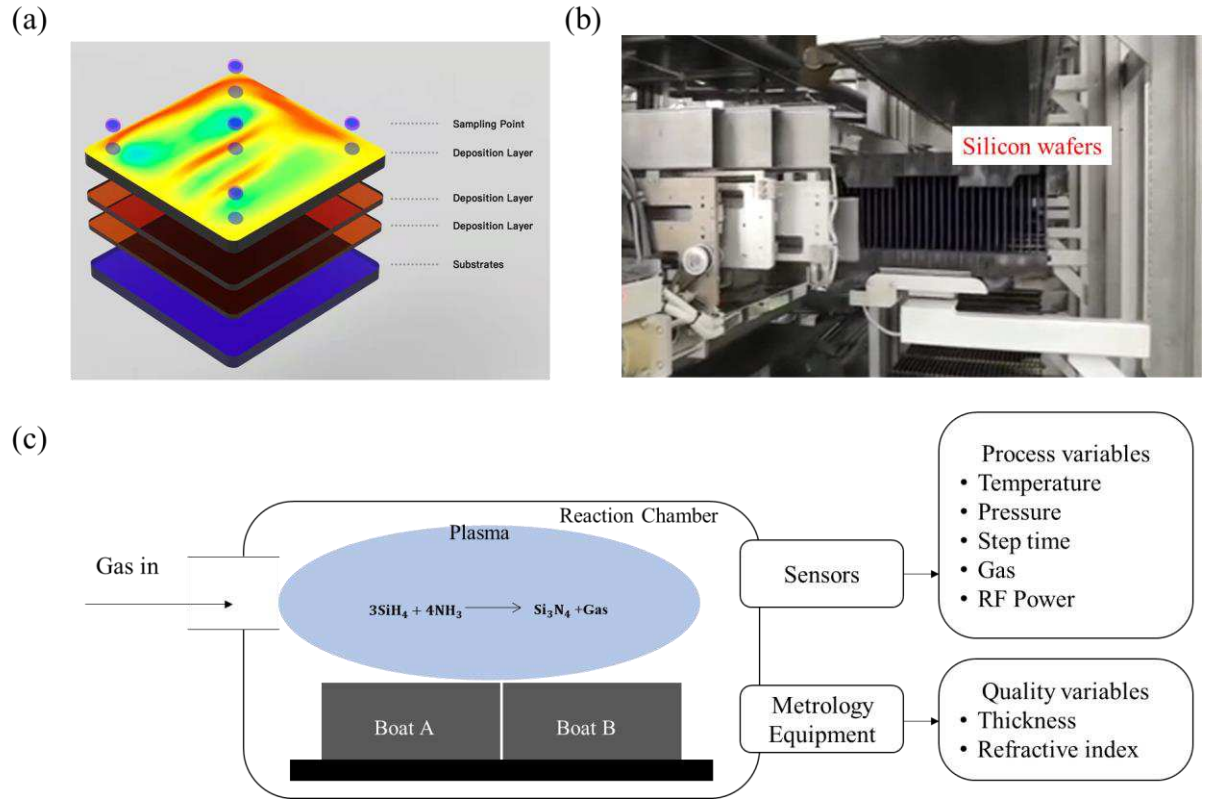

Fig. 2 The experimental setup. (a) The digital twin visualization of the PECVD processing. (b) The real manufacturing process of silicon wafers using the PECVD technology. (c) The diagram of the silicon solar-cell coating with the PECVD processing.

The total collected dataset includes 800 pairs of processing parameters and corresponding quality parameters of a thickness (TN) and refractive index (RI). The processing parameters were sampled by real-time sensors, which include RF power, temperature, pressure, gas flow, and deposition time of different processing steps. The median of parameters for each processing step is selected and reorganized into vectors of 60 parameters. The quality variables including film thickness and refractive index were collected and measured from two different boats. Then all the quality variables and processing data were associated through the equipment information. All the associated data has been stripped of outliers before the model training.

4.2. Result

The proposed MTGPVM model and learning algorithm are evaluated in this section using the above PECVD real metrology data. Different basic kernel functions are compared for the covariance function of MTGPVM. After that, the active learning of MTGPVM model using different query functions is compared.

4.2.1. Covariance functions

The proposed MTGPVM model mainly includes two kernel functions k_f and k_h which need to be confirmed in the engineering application. The quality data collected from the average thickness and average refractive index of boats are utilized to evaluate different combinations of basic kernel functions. In this study, the alternatives k_f are the SE kernel and the Ma3 kernel while the choices k_h are the bias kernel and the linear kernel. The MTGPVM model is trained using a natural gradient without active learning for this part. 500 training data and 200 testing data are randomly chosen from the original dataset and the quality variables are generated from four quality factors of two boats. After natural gradient for optimization, the Adam optimizer with 1000 iteration times is utilized for further optimization. The corresponding results are demonstrated in Table 1. The error of the combination of Ma3 kernel and the linear kernel is the smallest in both the training set and the test set. The combination of SE kernel and linear kernel takes second place. The MAPE is in the range of 0.99%-1.56% for predicted thickness and in the range of 0.76%-0.84% for predicted refractive index. Linear-based kernel functions lead to better performance compared with bias-based kernel functions. The biggest error among four kernel alternatives is from the combination of SE kernel and bias

kernel and results in 1.33%-1.79% MAPE for thickness and 2.16%-2.55% MAPE for refractive index.

Table 1 MAPE of quality prediction with different kernel alternatives.

Kernel Alternatives	Boat A				Boat B			
	Thickness		Refractive Index		Thickness		Refractive Index	
	Train	Test	Train	Test	Train	Test	Train	Test
Ma3+Linear	**0.0099**	**0.0146**	**0.0076**	**0.0079**	**0.0104**	**0.0156**	**0.0086**	**0.0084**
SE+Linear	0.0105	0.0136	0.0157	0.0159	0.0129	0.0159	0.0126	0.0126
Ma3+Bias	0.0124	0.0145	0.0163	0.0165	0.0167	0.0167	0.0160	0.0157
SE+Bias	0.0144	0.0133	0.0216	0.0217	0.0179	0.0171	0.0253	0.0255

4.2.2. Sampling criterions

The performance of three informative sampling methods and adaptive weight coefficients is examined in this part. The covariance function of MTGPVM is based on Ma3 and linear kernel. The number of the initial labeled training set is 30. Each query selects the most informative 10 unlabeled data. The validation set includes 100 labeled data. The MAPE of the four predicted quality factors using the validation set are recorded in the active learning process.

Fig. 3 (a)-(d) show the results of weighting coefficients considering only the film thickness or refractive index. It can be seen from these figures that the distribution characteristics of the original data set make the sampling results of the uncertain query and the combined query methods the same. The diversity information based on film thickness causes the four errors to decrease most rapidly at the beginning.

Fig. 3(e)-(h) demonstrate the active learning results with non-adaptive weight coefficients and the weights of thickness and refractive index are uniform and fixed. The diversity sampling-based learning has the lowest error with 30-50 labeled data. The combined active learning only has a lower error for the quality data of Boat A. Therefore, the weight coefficients of diversity sampling should be bigger than that of the uncertainty sampling initially.

Fig. 3(i)-(l) illustrate the results of adaptive active learning. The initial weight is set to 0. The learning rate for adaptive weight coefficients is 0.03. In this case, the combination

sampling can exploit the diverse information at the beginning. Although there is almost no difference between the UNs and AUNs, the adaptive active learning methods based on diverse information and combined information are accelerated compared with DIs and ACOs. After training with 150 labeled samples, all adaptive active learning methods bring about 2% MAPE for thickness and 0.5% MAPE for refractive index. Therefore, the learning speed based on the proposed active learning method is increased compared with gradient descent using vast labeled data.

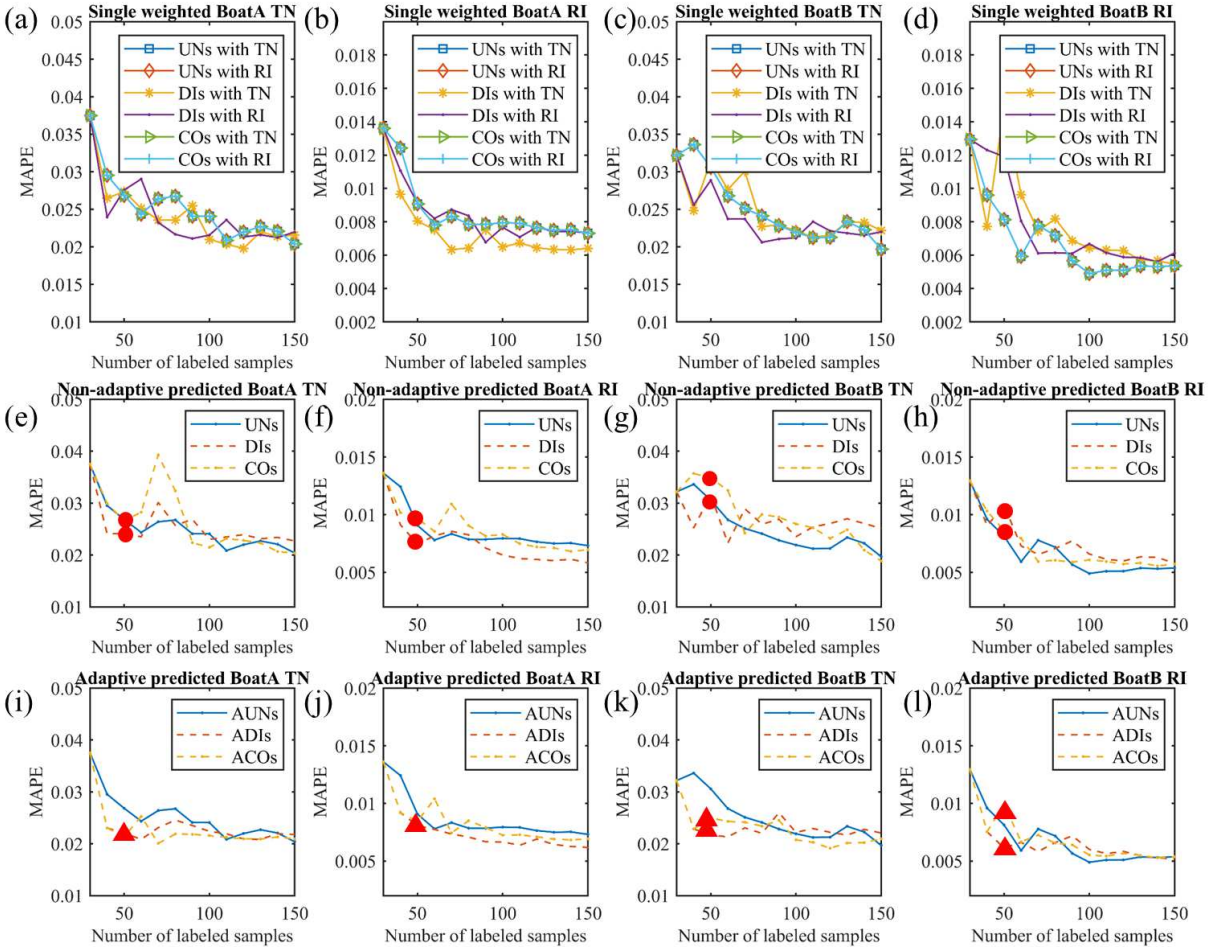

Fig. 3 The active learning results using different sampling criteria.

5. Comparison and discussion

The comparison results with other multitask learning methods are demonstrated in Table 2. The labeled training set has 500 data and the test set has 200 data. The proposed method based on multitask Gaussian process is evaluated with the best performance in both the training set and test set. Although the SVR method with ChiSquare kernel has a lower error compared to

the training set, its error on the test set is more than twice that on the training set. Similarly, the overfitting has also happened to the SVR with RBF kernel. Additionally, the random forest ensemble (RFE) is also compared with MTGPVM and obtains worse performance in the test set. Besides, we built a multilayer perceptron (MLP) model to compare our method with neural networks. The MLP is not qualified for this case study. Therefore, the proposed method has a more robust performance and can avoid overfitting compared with another method.

Table 2 MAPE with different multitask learning models.

Methods	Kernel	Boat A				Boat B			
		Thickness		Refractive Index		Thickness		Refractive Index	
		Train	Test	Train	Test	Train	Test	Train	Test
MTGPVM	**Ma3+Linear**	**0.0135**	**0.0132**	**0.0041**	**0.0043**	**0.0173**	**0.0180**	**0.0051**	**0.0055**
SVR	ChiSquare	0.0122	0.0395	0.0033	0.0143	0.0146	0.0334	0.0039	0.0134
SVR	RBF	0.1850	0.9364	0.1852	0.9364	0.1854	0.9373	0.1852	0.9355
RFE	-	0.0133	0.0160	0.0037	0.0044	0.0145	0.0208	0.0044	0.0056
MLP	-	0.2148	0.5697	0.1674	0.2045	0.2577	0.5934	0.1651	0.1757

The active learning process with SVR, RFE, and proposed MTGPVM is illustrated in Fig. 4. The sampling method is based on the diversity information of (20) without adaptive modulation of weight coefficients. Fig. 4(a)-(d) are the predicted results of four quality variables. Although the RFE based active learning method has a stable variation with relatively smaller errors, the errors of MTGPVM based method decreases more rapidly. The errors of MTGPVM are lower than RFE in Fig. 4(a) and (c), which means the proposed MTGPVM model is more suitable for the thickness prediction. In Fig. 4(b), the SVR with the ChiSquare kernel gets a lower error initially in comparison to MTGPVM model, but the rate of error decline is relatively smaller.

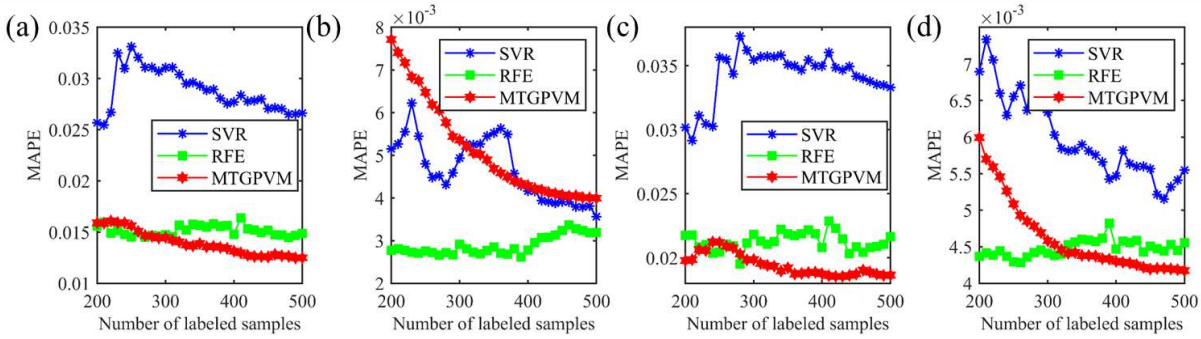

Fig. 4 The active learning results with SVR, RBF and the proposed MTGPVM model. Four quality variables are composed of: (a) Thickness of Boat A; (b) Refractive index of Boat A; (c) Thickness of Boat B; (d) Refractive index of Boat B.

In conclusion, the MTGPVM with suitable basic kernels may lead to better performance than other multitask learning methods, such as random forest ensemble and neural networks. With the assistance of the proposed active learning method, the MTGPVM method can outperform other methods as the number of labeled samples increases.

6. Conclusion

Virtual metrology plays an important role in the controller of semiconductor manufacturing. It usually learns from real metrology and predicts quality variables using processing data. The proposed methods solve the problems of multivariable prediction and learning with a small labeled sample size by multitask Gaussian processes and adaptive active learning. The actual engineering data from the chemical vapor deposition process within two boats are utilized to evaluate the effectiveness of the proposed method. Firstly, different kernel alternatives are compared and the combination of Matern kernel and linear kernel leads to the 0.99%-1.56% MAPE of thickness and 0.76%-0.84% MAPE of refractive index. Subsequently, the proposed adaptive active learning algorithm for virtual metrology results in ~0.5% MAPE of refractive index with 150 labeled training data. Ultimately, compared with other multitask learning methods, the proposed MTGPVM model consistently performs with lower error regardless of the training set or the test set. Furthermore, the MTGPVM can outperform other methods with the assistance of the proposed active learning method.

The future direction of research can be the improved combination of different advanced sampling methods for the active learning and the ensemble of different multitask learning models. Moreover, the improved VM model can be combined with the digital twin model to be applied to other intelligent manufacturing fields in the future.

Statements and Declarations

Funding

This study was supported by the National Natural Science Foundation of China with No. 51775108. This study was also supported by "State Key Laboratory of Transducer Technology" with No. SKT2102.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Shanling Ji, Zhisheng Zhang and Zhijie Xia. The first draft of the manuscript was written by Shanling Ji and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. Khan A, Moyne J,Tilbury D (2007) An Approach for Factory-Wide Control Utilizing Virtual Metrology. IEEE Transactions on Semiconductor Manufacturing 20(4): 364-375. https://doi.org/10.1109/TSM.2007.907609
2. Hsieh Y, Wang T, Lin C et al (2021) Convolutional Neural Networks for Automatic Virtual Metrology. IEEE Robotics and Automation Letters 6(3): 5720-5727. https://doi.org/10.1109/LRA.2021.3084882
3. Wu X, Chen J, Xie L et al (2021) Convolutional Neural Networks for Multi-Stage Semiconductor Processes. Journal of Chemical Engineering of Japan 54449-455. https://doi.org/10.1252/jcej.20we139
4. Lee K,Kim C (2020) Recurrent feature-incorporated convolutional neural network for virtual metrology of the chemical mechanical planarization process. Journal of Intelligent Manufacturing 31(1): 73-86. https://doi.org/10.1007/s10845-018-1437-4
5. Kang P, Kim D,Cho S (2016) Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing. Expert Systems with Applications 5185-106. https://doi.org/https://doi.org/10.1016/j.eswa.2015.12.027

6. Di Y, Jia X,Lee J (2017) Enhanced Virtual Metrology on Chemical Mechanical Planarization Process using an Integrated Model and Data-Driven Approach. International Journal of Prognostics and Health Management 8https://doi.org/10.36001/ijphm.2017.v8i2.2641

7. Lee S-k, Kang P,Cho S (2014) Probabilistic local reconstruction for k-NN regression and its application to virtual metrology in semiconductor manufacturing. Neurocomputing 131427-439. https://doi.org/https://doi.org/10.1016/j.neucom.2013.10.001

8. Wan J,McLoone S (2018) Gaussian Process Regression for Virtual Metrology-Enabled Run-to-Run Control in Semiconductor Manufacturing. IEEE Transactions on Semiconductor Manufacturing 3112-21. https://doi.org/10.1109/TSM.2017.2768241

9. Bonilla E, Chai K,Williams C (2008) Multi-Task Gaussian process prediction. Proc. Adv. Neural Inf. Process. Syst 20153-160.

10. Álvarez M, Rosasco L,Lawrence N (2012) Kernels for Vector-Valued Functions: A Review. Foundations and Trends® in Machine Learning 4195-266. https://doi.org/10.1561/2200000036

11. Park C, Kim Y, Park Y et al (2018) Multitask learning for virtual metrology in semiconductor manufacturing systems. Computers & Industrial Engineering 123209-219. https://doi.org/https://doi.org/10.1016/j.cie.2018.06.024

12. Shim J,Kang S (2022) Domain-adaptive active learning for cost-effective virtual metrology modeling. Computers in Industry 135103572. https://doi.org/https://doi.org/10.1016/j.compind.2021.103572.

13. Rasmussen C,Williams C (2006) Gaussian Processes for Machine Learning.MIT Press.

14. Park C, Borth D, Wilson N et al (2021) Robust Gaussian process regression with a bias model. Pattern Recognition 124108444. https://doi.org/https://doi.org/10.1016/j.patcog.2021.108444

15. Dürichen R, Pimentel M, Clifton L et al (2015) Multitask Gaussian Processes for Multivariate Physiological Time-Series Analysis. IEEE Transactions on Biomedical Engineering 62(1): 314-322. https://doi.org/10.1109/TBME.2014.2351376

16. Cai H, Feng J, Yang Q et al (2020) A virtual metrology method with prediction uncertainty based on Gaussian process for chemical mechanical planarization. Computers in Industry 119103228. https://doi.org/https://doi.org/10.1016/j.compind.2020.103228

17. Wu D,Huang J (2022) Affect Estimation in 3D Space Using Multi-Task Active Learning for Regression. IEEE Transactions on Affective Computing 13(1): 16-27. https://doi.org/10.1109/TAFFC.2019.2916040

18. Opper M,Archambeau C (2009) The Variational Gaussian Approximation Revisited. Neural Computation 21(3): 786-792. https://doi.org/10.1162/neco.2008.08-07-592

19. Matthews A, Wilk M, Nickson T et al (2016) GPflow: A Gaussian Process Library using TensorFlow. Journal of Machine Learning Research 18