

## Research Article

**Cite this article:** Song Z, Zhao W, Zhang X, Ke M, Fang W and Lyu B (2024). Stacking ensemble learning based material removal rate prediction model for CMP process of semiconductor wafer. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **38**, e17, 1–17  
<https://doi.org/10.1017/S0890060424000167>

Received: 09 December 2023  
Revised: 29 June 2024  
Accepted: 01 July 2024


### Keywords:

chemical mechanical polishing; material removal rate; ensemble learning; semiconductor wafer

### Corresponding author:

Lyu Binghai;  
Email: [icewater7812@126.com](mailto:icewater7812@126.com)

# Stacking ensemble learning based material removal rate prediction model for CMP process of semiconductor wafer

Zhulong Song<sup>1,2</sup> , Wenhong Zhao<sup>1,2</sup>, Xiao Zhang<sup>1,2</sup>, Mingfeng Ke<sup>1,2</sup>, Wei Fang<sup>1,2</sup> and Binghai Lyu<sup>1,2</sup>

<sup>1</sup>College of Mechanical Engineering, Zhejiang University of Technology, Hangzhou, China and <sup>2</sup>Ultra-Precision Machining Center, Key Laboratory of Special Purpose Equipment and Advanced Processing Technology, Ministry of Education and Zhejiang Province, Zhejiang University of Technology, Hangzhou, China

## Abstract

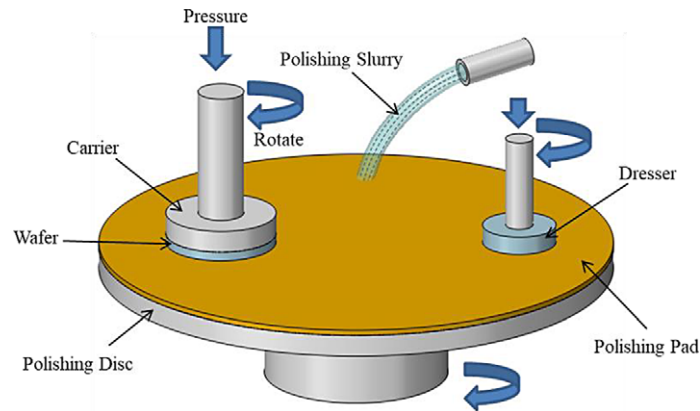
The material removal rate (MRR) serves as a crucial indicator in the chemical mechanical polishing (CMP) process of semiconductor wafers. Currently, the mainstream method to ascertain the MRR through offline measurements proves time inefficient and struggles to represent process variability accurately. An efficient MRR prediction model based on stacking ensemble learning that integrates models with disparate architectures was proposed in this study. First, the processing signals collected during wafer polishing, as available in the PHM2016 dataset, were analyzed and preprocessed to extract statistical and neighbor domain features. Subsequently, Pearson correlation coefficient analysis (PCCA) and principal component analysis (PCA) were employed to fuse the extracted features. Ultimately, random forest (RF), light gradient boosting machine (LightGBM), and backpropagation neural network (BPNN) with hyperparameters optimized by the Bayesian Optimization Algorithm were integrated to establish an MRR prediction model based on stacking ensemble learning. The developed model was verified on the PHM2016 benchmark test set, and a Mean Square Error (MSE) of 7.72 and a coefficient of determination ( $R^2$ ) of 95.82% were achieved. This indicates that the stacking ensemble learning based model, integrated with base models of disparate architectures, offers considerable potential for real-time MRR prediction in the CMP process of semiconductor wafers.

## Introduction

Chemical mechanical polishing (CMP) is the predominant method for semiconductor wafer processing. A representative process schematic is depicted in [Figure 1](#). Material removal during CMP predominantly occurs via the synergistic actions of chemistry and machinery among the wafer, polishing slurry, and polishing pad (Zhang et al., 2021). Real-time monitoring of the Material Removal Rate (MRR) during the polishing phase provides immediate insights into the processing status, delivering crucial information for subsequent applications, including electrical characterization and layout design. However, offline measurement using precision instruments is currently the main method to obtain MRR (Lee, 2019), which falls short of facilitating real-time monitoring. In addition, measurement precision may be compromised by operator-induced variability. Consequently, there is a pressing need for a precise and efficient MRR prediction model.

The Preston equation (Evans et al., 2003):  $MRR = K_p P^\alpha V^\beta$  is often used to construct the MRR model, where  $P$  denotes the downward pressure applied to the wafer,  $V$  denotes the relative rotating speed between the wafer and the polishing pad,  $K_p$  is the Preston coefficient, and  $\alpha$  and  $\beta$  are the parameters depending on the operating conditions. However, due to the complexity of the CMP process, the model construction method based on the Preston equation is difficult to consider various process variables more comprehensively to accurately predict the MRR. Coinciding with advancements in machine learning and deep learning, data-driven regression models for MRR prediction have emerged (Wang et al., 2017; Li et al., 2018; Xu et al., 2021). The relative studies show that the integrated model performs better than the individual models therein (Di et al., 2021; Li et al., 2019), which focused on crafting MRR prediction models through ensemble learning and sought to integrate models of similar architectures or principles. Yet, the distinctive strength of data-driven model integration lies in its aptitude to assimilate samples from varied perspectives when the base models exhibit substantial divergence in structure or principle (Wolpert David, 1992).

The main contribution of our work is that an MRR prediction model with high precision for CMP process has been built. By means of feature extraction first and then feature fusion, enabling the model to learn rich data features on the foundation of maintaining a low number of parameters, so as to be able to achieve high-precision real-time prediction. Specifically, it's the



**Figure 1.** Schematic diagram of a typical CMP process.

first attempt to construct the MRR prediction model by integrating base models with substantial disparities, and the effectiveness of which was corroborated on the public 2016th Prognostics and Health Management dataset (Jia *et al.*, 2021) (PHM2016).

### Related work

Existing MRR predictive models for CMP principally bifurcate into two categories: those grounded in mechanical and chemical principles, and those deriving from data-driven methodologies. Zhao and Chang (2002) developed a closed-loop MRR prediction model by studying elastoplastic microcontact mechanics and polishing pad wear theory. Experimental results underscore the correlation between MRR and factors such as abrasive concentration and abrasive radius in the polishing slurry. Xu *et al.* (2020) constructed a CMP analytical model predicated on the governing equation of plate theory, chemical reaction kinetics, and wear theory, and the influence of variables including pad elastic modulus, temperature distribution, carrier rotation speed, and so forth on MRR have been explored. These scholarly endeavors invested in unraveling the relationships between diverse process parameters and MRR, aimed at formulating a theoretical model bridging process parameters and MRR. Despite the variety of wafer and slurry materials, the numerous related process parameters, and the complex conditions found in the CMP process, finding a comprehensive theory that clearly explains the inherent material removal mechanisms is still a challenge. Consequently, approaches relying on physical and chemical principles for MRR prediction are inherently constrained.

A machine learning model, in theory, retains the capability to approximate an arbitrarily complex mathematical landscape (Hanin and Rolnick, 2019). Through successive refinements in the course of optimization, the model can autonomously deduce the inherent relationships between process parameters and MRR, bypassing the requirement for strenuous theoretical analysis and computational complexities. Xu *et al.* (2021) proposed a data-driven neural network (NN) model based on CMP experiments to predict MRR and investigated the influence of the oxidizer concentration and the inhibitor concentration, as well as the chelating agent concentration and the surfactant concentration on the prediction of MRR. Li *et al.* (2018) utilized random forest (RF) to predict MRR through discriminating between fine and coarse polishing. Wang *et al.* (2017) devised an optimized Deep Belief Network (DBN) to investigate the relationship between MRR and polishing operation parameters such as pressure and rotational

speeds of the wafer and pad. Furthermore, the strength of the data-driven approach in constructing MRR prediction models resides in the potential enhancement of prediction precision through integration of multiple models trained with identical samples. Di *et al.* (2021) put forth an ensemble learning based model, incorporating k-nearest neighbor (KNN), support vector machine (SVM) and logistic regression (LR) to predict the MRR, and proved the validity of ensemble learning through experiments.

High-quality polishing process datasets like PHM2016 are difficult to obtain, which has prompted many scholars to conduct research using this data. For instance, Li *et al.* (2019) employed a stacking ensemble learning method based on classification and regression trees (CART) and the extreme learning machine (ELM). This method had a Root Mean Squared Error (RMSE) of 4.64 on the test dataset, which far exceeds the accuracy of models based solely on the Preston equation, or single models like RF, GBT, and ERT. Zhang *et al.* (2021) used the residual convolutional neural network (ResCNN) to build an MRR prediction model for the CMP process and achieved a Mean Square Error (MSE) of 6.72 on the test set. Their experimental results highlight the impact that the usage quantity of each consumable in the CMP process has on MRR. However, utilizing convolutional neural networks (CNNs) demands the conversion of input attributes into two or three-dimensional matrices, putting certain requirements on the length of the input features. Additionally, the sequential computation process of CNN models could impact real-time online predictions. Thus, the use of CNN-based prediction models may be limited due to these factors.

Therefore, the goal of this study is to construct an MRR prediction model of CMP based on ensemble learning, which fuses the neural network models and tree models with significantly different structures and principles, and ultimately obtains higher prediction accuracy on the test set compared to other methods based on ensemble learning.

### Stacking ensemble learning based predictive modeling

#### Algorithm framework

The framework of the proposed algorithm is depicted in Figure 2. Initially, the dataset was partitioned into training and testing subsets, each serving the dual purposes of model training and performance evaluation, respectively. In the training phase, noise samples within the training set were preemptively discarded. Subsequently, feature engineering techniques were employed to transform the raw input data into representative features. Finally, a stacking ensemble

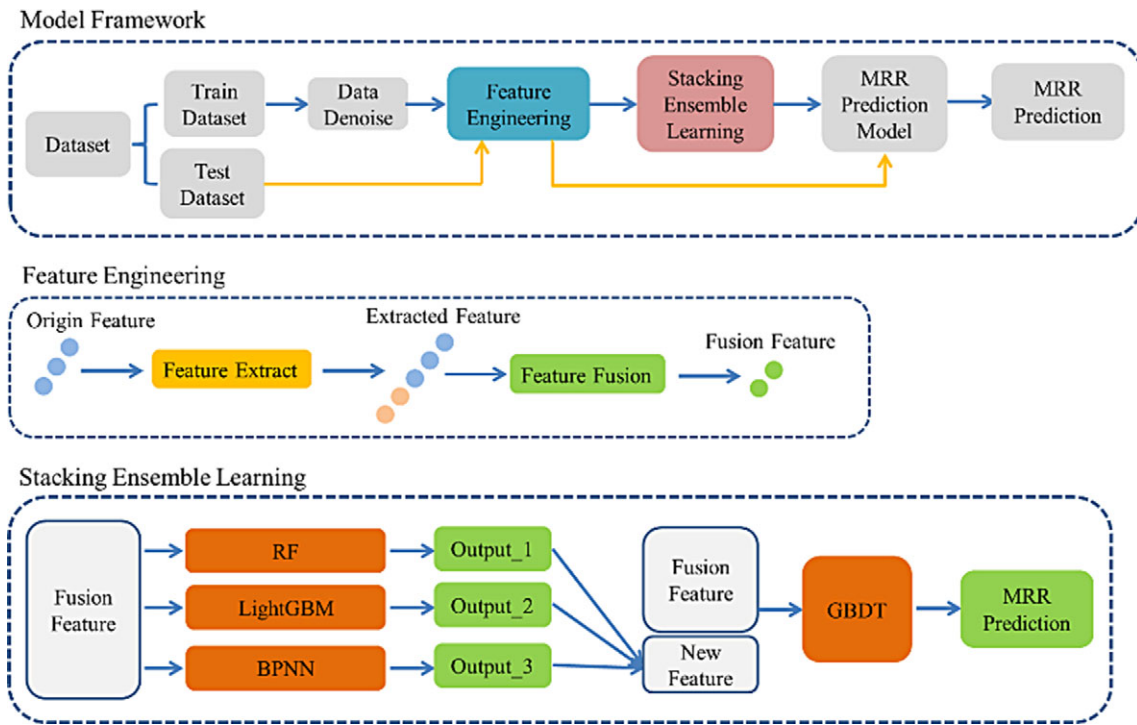


Figure 2. Schematic diagram of algorithm framework.

approach (Tang et al., 2022) integrating RF (Breiman, 2001), light gradient boosting machine (LightGBM) (Ke et al., 2017), and back-propagation neural network (BPNN) (Rumelhart et al., 1986) was executed to build the MRR prediction model. In the prediction phase, the test samples were subjected to identical feature engineering procedures as in the training process, and the trained model was then harnessed to forecast the MRR.

### Feature fusion

Potential correlations might exist among the monitored attributes during the CMP process, leading to feature redundancy. This unnecessary repetition interferes with model training by adding extra information, which takes attention away from key features and reduces accuracy (Batista et al., 2004). Moreover, this increases the input dimensionality, substantially escalating computational demands and prediction latency, consequently restricting the feasibility of real-time prediction.

Pearson correlation coefficient analysis (PCCA) (Malik et al., 2021) was utilized to analyze the correlations among features. Features with strong correlation were considered as a singular group. Principal component analysis (PCA) (Pearson and Karl, 2010) was subsequently implemented to fuse features within the same group and diminish dimensionality. Ultimately, to assess the efficacy of feature fusion, a comparison was made between the accuracy and efficiency of MRR prediction prior to and subsequent to feature fusion.

### Pearson correlation coefficient analysis

PCC quantifies the linear correlation between variables  $X$  and  $Y$ , ranging from  $-1$  to  $1$ . Equation (1) details the PCC calculation:

$$\rho_{X,Y} = \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}} \quad (1)$$

where  $N$  represents the sample size,  $X$  and  $Y$  denote two distinct features of the sample. While covariance does indicate correlations between random variables (positive when covariance  $>0$ , negative when covariance  $<0$ ), its magnitude is heavily influenced by the variances of  $X$  and  $Y$ , hence prohibiting the deduction of correlations between variables solely from covariance. Nevertheless, the PCC provides a precise portrayal of correlations between variables, independent of dimensional disparities between  $X$  and  $Y$ .

### Principal component analysis

The core principle of PCA pertains to the transformation of original multi-dimensional features into orthogonal principal components. These components primarily hold the valuable information from the original features, while lessening repetition, therefore simplifying the complexity of the original feature space. Given a dataset  $X$  comprising  $n$  samples with  $m$  features each, shown by equation (2),  $X_m(n)$  represents the  $m$ -th eigenvalue of the  $n$ -th sample. The PCA process can be outlined as follows:

$$X_{n \times m} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_m(1) \\ x_1(2) & x_2(2) & \cdots & x_m(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(n) & x_2(n) & \cdots & x_m(n) \end{bmatrix} \quad (2)$$

(1) Calculating the covariance matrix  $P$  of normalized  $X$ , which can be employed to describe the correlations among  $m$  variables in the dataset, according to equation (3), and getting the eigenvalue  $\lambda$  and eigenvector  $E$  of  $P$  in line with equation (4),  $D$  represents the diagonal matrix.

$$P = \text{cov}(X) = \frac{1}{m-1} X^T X \quad (3)$$

$$P = EDE^T \quad (4)$$

(2) The cumulative contribution of the initial  $k$  principal components, symbolized as  $a_k$ , is calculated as equation (5). When  $a_k$  surpasses the pre-established cumulative contribution threshold  $T$ , the  $k$  principal components become viable replacements for the original  $m$  features. In other words, the PCA process contracts the feature dimensionality from  $m$  to  $k$ .

$$a_k = \frac{\lambda_k}{\sum_{i=1}^m \lambda_i} \tag{5}$$

**Prediction model**

RF, LightGBM, and BPNN models were integrated to construct the MRR prediction model. In terms of model architectures, RF and LightGBM can be conceptualized as tree-based statistical models, whereas BPNN is a neural network model. Owing to the discrepancies in their model structures, they are capable of extracting distinct information from the same set of samples (Pearson and Karl, 2010).

**Random forest**

RF, a bagging-based regression prediction model, trains each decision tree on independently sampled subsets, as depicted in Figure 3, subsequently aggregating the outputs of all trees to procure the final model output.

In practical applications, the accuracy of individual decision trees within an RF can vary. Relying on simple averaging may lead to a drop in the overall accuracy of the RF model due to the influence of lower-accuracy decision trees. Consequently, a weighted ensemble method was utilized during RF training, wherein the significance of each decision tree was determined based on its MSE. The outputs of the decision trees were subsequently weighted to compute the final output of the RF model, as depicted in equation (6).

$$f(x) = \sum_{i=1}^n \alpha_i * T_i(x) \tag{6}$$

where  $\alpha_i$  and  $T_i$  represents the significance and predicted value of the  $i$ -th decision tree.

**LightGBM**

LightGBM is an enhanced algorithm based on gradient boosting decision trees (GBDT) (Friedman, 2001), boasting superior computational efficiency and lower memory demands, thus making it

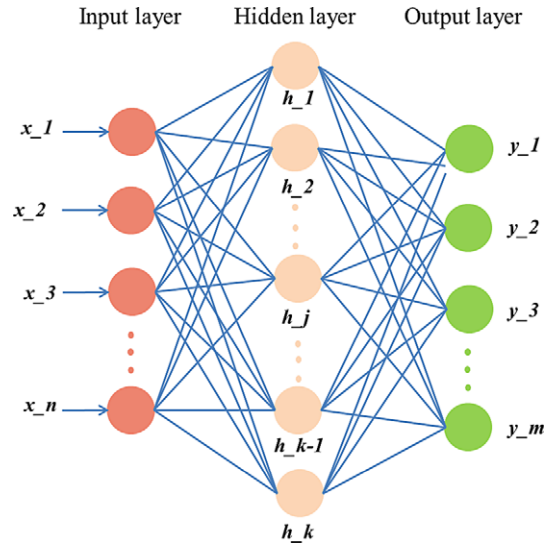


Figure 5. Schematic diagram of BPNN.

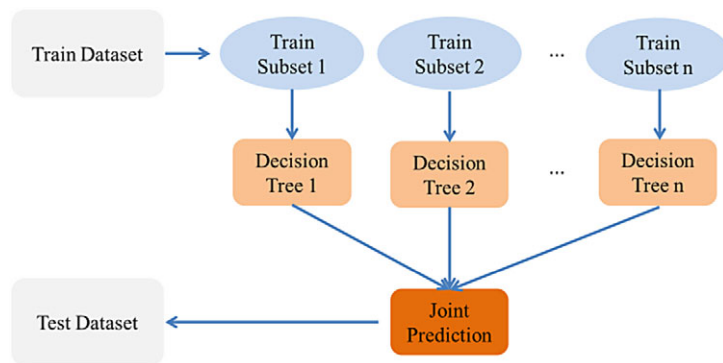


Figure 3. Schematic diagram of random forest.

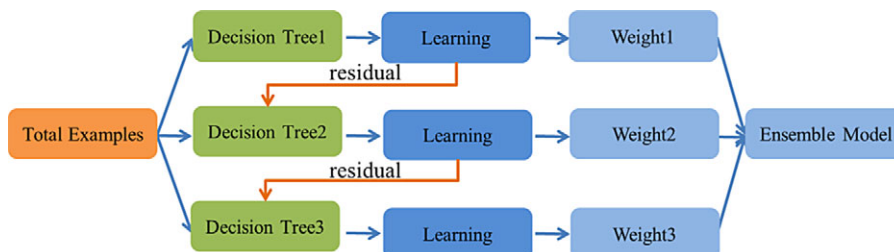


Figure 4. Schematic diagram of LightGBM.

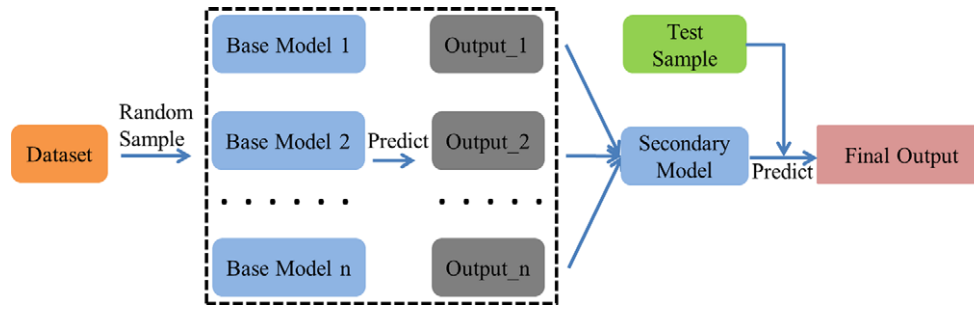


Figure 6. Schematic diagram of stacking ensemble learning.

Table 1. Experimental environment

| Software/package | Version      | Description             |
|------------------|--------------|-------------------------|
| Pycharm          | 2020.2.3_x64 | IDE for Python          |
| Python           | 3.6.6        | Programming language    |
| Sklearn          | 0.24.2       | Machine learning lib    |
| Numpy            | 1.19.5       | Numerical computing lib |
| Pandas           | 1.1.5        | Data manipulation lib   |
| Matplotlib       | 3.3.4        | Data visualization lib  |

ideal for high-performance prediction applications. Much like RF, GBDT also constitutes a tree-based statistical model, but its trees predict the residual or difference between the estimated and actual values from all preceding trees. Figure 4 depicts the LightGBM training process.

For a sample  $x$ , the prediction process can be described by equation (7):

$$f(x) = f_0(x) + \sum_{t=1}^T \sum_{j=1}^J c_{tj} I(x \in R_{tj}) \quad (7)$$

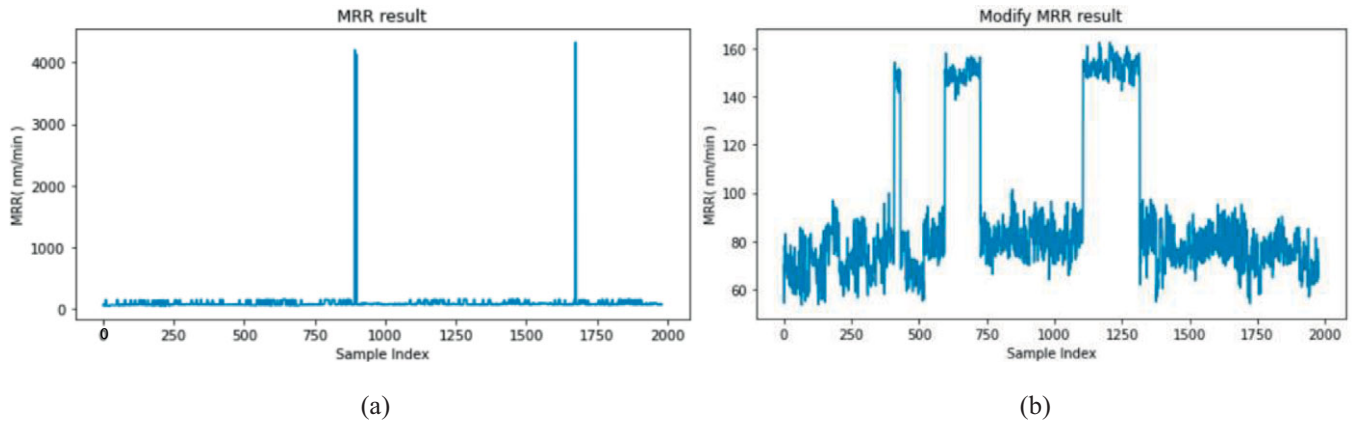
where  $T$  denotes the number of CART decision trees (Zounemat-Kermani et al., 2020),  $J$  signifies the number of leaf node regions in each tree, and  $c_{tj}$  stands for the predicted value of the  $j$ -th leaf node region on the  $t$ -th CART tree. When  $x$  falls within the set  $R_{tj}$ ,  $I$  equals to 1; otherwise, it equals to 0.

### BP neural network

The BPNN (Ruan, 2021) is a foundational deep learning model. Its robust nonlinear function approximation ability can theoretically fit any complex mathematical function. Figure 5 illustrates a three-layer BPNN architecture consisting of input, hidden, and output layers. The neurons in the input layer correspond to independent variables, such as the rotation speed of the polishing pad or temperature. Meanwhile, the neurons in the output layer represent dependent variables, for instance, the MRR. The neuron count in the hidden layer, which determines the model complexity, is user-defined. Typically, this range is approximately estimated using empirical equation (8) (Yu et al., 2022), which is based on practical observations and experiences in the field of neural network

Table 2. Process variables during the CMP process

| Number | Signal                       | Description                        |
|--------|------------------------------|------------------------------------|
| x1     | MACHINE_ID                   | Machine ID                         |
| x2     | MACHINE_DATA                 | Wafer ring location ID             |
| x3     | TIMESTAMP                    | Time                               |
| x4     | WAFER_ID                     | Wafer ID                           |
| x5     | STAGE                        | Stage ID                           |
| x6     | CHAMBER                      | Chamber ID                         |
| x7     | USAGE_OF_BACKING_FILM        | The polish-pad backing film usage  |
| x8     | USAGE_OF_DRESSER             | The dresser usage                  |
| x9     | USAGE_OF_POLISHING_TABLE     | The polishing table usage          |
| x10    | USAGE_OF_DRESSER_TABLE       | The dresser table usage            |
| x11    | PRESSURIZED_CHAMBER_PRESSURE | The pressure of chamber            |
| x12    | MAIN_OUTER_AIR_BAG_PRESSURE  | The pressure of main outer air bag |
| x13    | CENTER_AIR_BAG_PRESSURE      | The pressure of center air bag     |
| x14    | RETAINER_RING_PRESSURE       | The pressure of retainer ring      |
| x15    | RIPPLE_AIR_BAG_PRESSURE      | The pressure of ripple airbag      |
| x16    | EDGE_AIR_BAG_PRESSURE        | The pressure of edge air bag       |
| x17    | USAGE_OF_MEMBRANE            | The polishing membrane usage       |
| x18    | USAGE_OF_PRESSURIZED_SHEET   | The wafer carrier sheet usage      |
| x19    | SLURRY_FLOW_LINE_A           | Flow rate of slurry type A         |
| x20    | SLURRY_FLOW_LINE_B           | Flow rate of slurry type B         |
| x21    | SLURRY_FLOW_LINE_C           | Flow rate of slurry type C         |
| x22    | WAFER_ROTATION               | Wafer rotating rate                |
| x23    | STAGE_ROTATION               | Stage rotating rate                |
| x24    | HEAD_ROTATION                | Head rotating rate                 |
| x25    | DRESSING_WATER_STATUS        | The dressing water status          |



**Figure 7.** MRR distribution of origin data and denoised data.

**Table 3.** Number of training and test set samples before and after denoise

| Datasets      | Origin | Denoised |
|---------------|--------|----------|
| Train Dataset | 1981   | 1777     |
| Test Dataset  | 424    | 391      |

**Table 4.** Result of dataset group

| Dataset Group ID | Process Stage | Chamber number | MRR       | Train samples | Test samples |
|------------------|---------------|----------------|-----------|---------------|--------------|
| I                | A             | 1,2,3          | 140 ~ 160 | 164           | 34           |
| II               | A             | 4,5,6          | 50 ~ 90   | 798           | 185          |
| III              | B             | 4,5,6          | 50 ~ 100  | 815           | 172          |

training. Where  $h$ ,  $r$ , and  $e$  denote the number of neurons in the input layer, hidden layer, and output layer, respectively, and 5 is a constant between 2 and 10.

$$h = \sqrt{r + e} + s \quad (8)$$

### Stacking ensemble learning

Since the models fused here have distinct architectures, stacking ensemble learning was selected to take advantage of their complementary strengths, thereby improving prediction

accuracy, robustness, and generalizability. Figure 6 illustrates the stacking procedure. The secondary model plays a key role in stacking, as it determines the optimal way to integrate and assign weights to the predictions of each base model based on their performance.

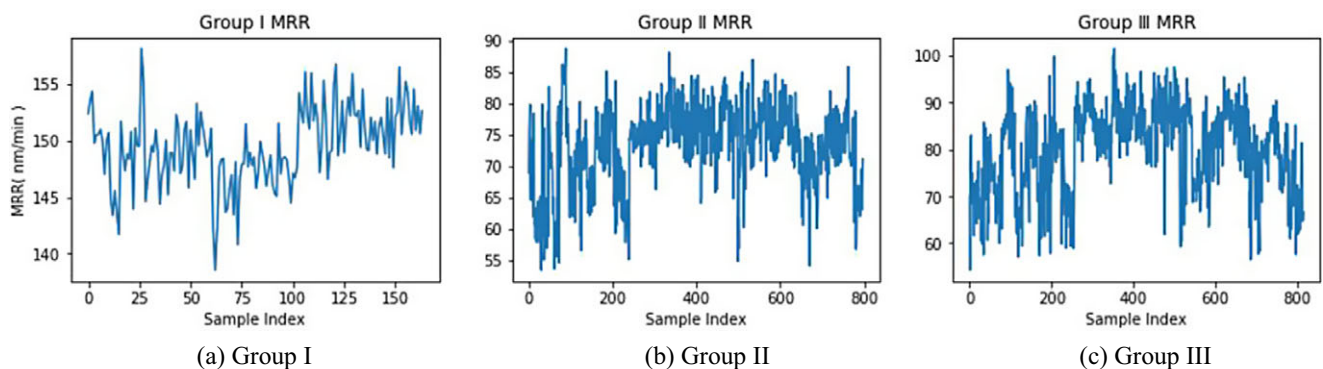
### Prediction evaluation indicators

Mean square error (MSE) (Köksoy, 2006) and correlation coefficient ( $R^2$ ) (Zhou et al., 2022) were utilized as evaluation metrics for the accuracy of MRR prediction model. As depicted in equation (9), MSE quantifies the discrepancy between the actual and predicted values of MRR. A smaller MSE value implies higher accuracy.  $R^2$ , as shown in equation (10), represents the correlation between the actual and predicted values of the MRR on the test dataset. A larger  $R^2$  value signifies greater accuracy.

$$MSE = \frac{1}{N} \sum_{i=1}^N (MRR_{pi} - MRR_{ti})^2 \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (MRR_{pi} - MRR_{ti})^2}{\sum_{i=1}^N \left( \frac{1}{N} \sum_{i=1}^N (MRR_{pi}) - MRR_{ti} \right)^2} \quad (10)$$

where  $MRR_{pi}$  and  $MRR_{ti}$  represent the predicted and actual values of MRR for the test samples, respectively, and  $N$  signifies the total number of test samples.



**Figure 8.** MRR distribution of each dataset group samples.

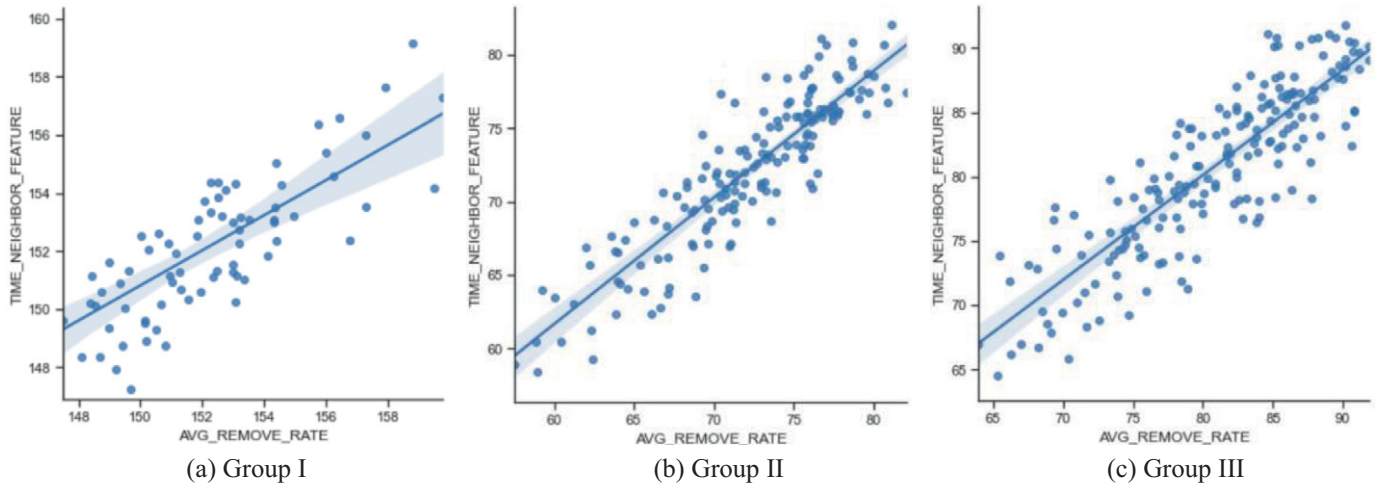


Figure 9. Correlation between time domain features and MRR.

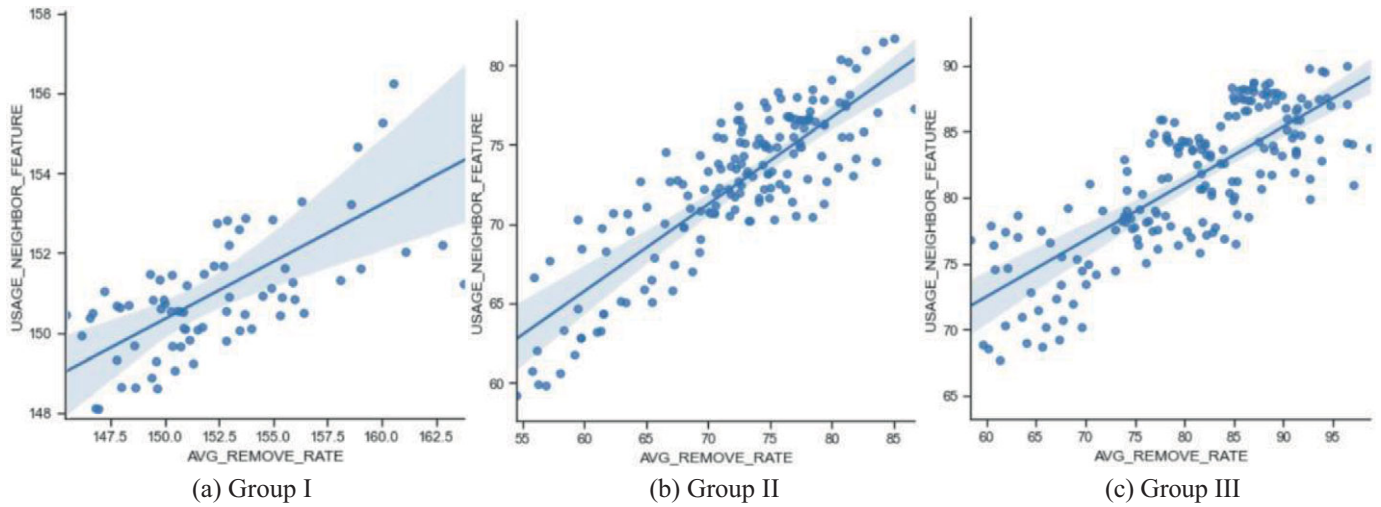


Figure 10. Correlation between usage domain features and MRR.

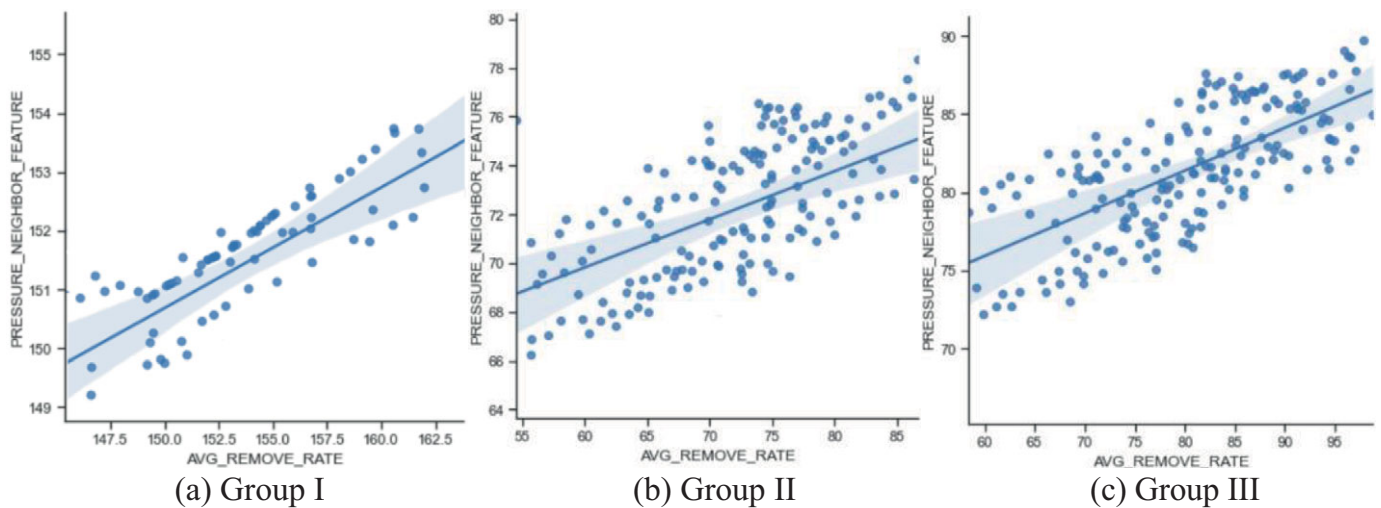
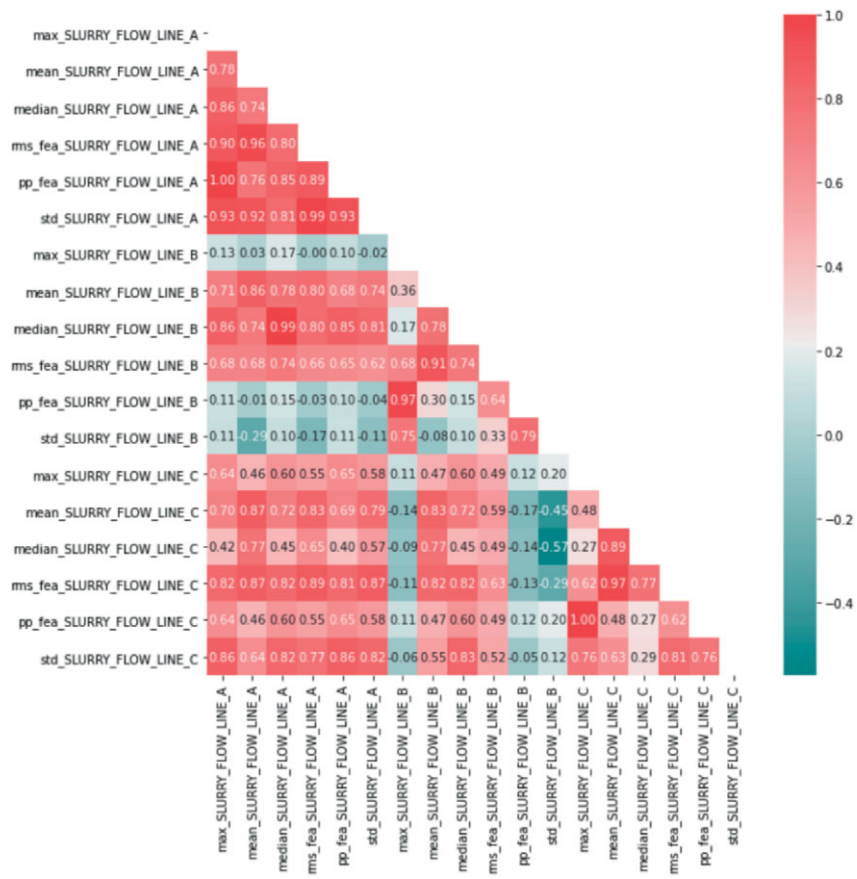
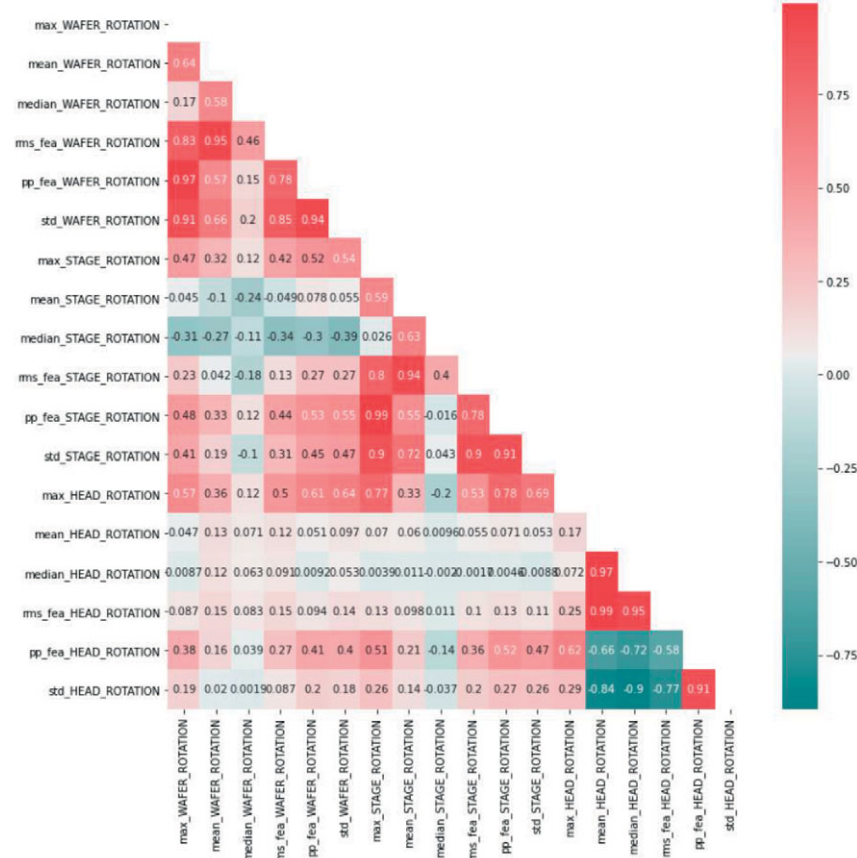


Figure 11. Correlation between pressure domain features and MRR.



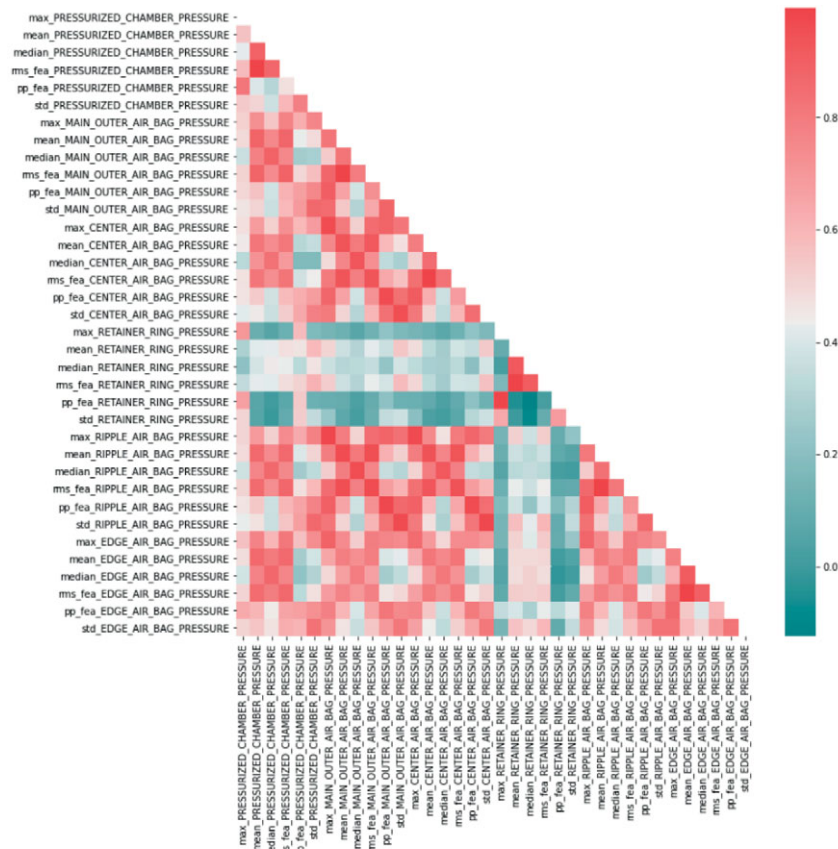
(a)



(b)

Figure 12. PCCA heatmap of (a) slurry flow rate, (b) speed, and (c) pressure.





(c)

Figure 12. Continued

## Case study

### Data analysis and preprocess

#### Dataset introduction

To validate the proposed MRR prediction model, experiments were conducted using the PHM Society’s 2016 open dataset for CMP MRR prediction on silicon wafers (Zhang et al., 2021). The dataset was divided into a training set (1981 samples) and a test set (424 samples). Each sample includes CMP processing signals and MRR measurements for a wafer at a given time point. Table 1 shows the main experimental environment and the corresponding software versions. Table 2 briefly outlines each signal variable.

#### Data denoise

In the wafer polishing experiments, each process variable was obtained from sensor measurements, which could potentially produce outlier values. Incorporating such anomalous samples into model training might undermine accuracy due to skewed data distributional (Sun et al., 2022). As depicted in Figure 7(a), the MRR of certain samples was significantly higher than that for others, indicating the presence of outliers. Figure 7(b) displays the MRR distribution following the exclusion of these abnormal samples.

In addition to outliers, samples with missing values, potentially due to sensor failures, cannot serve as valid training samples.

Table 3 presents a comparison of the total number of training and test set samples before and after noise removal.

#### Data split

As depicted in Figure 7, the MRR significantly ranges from 50 nm/min to 160 nm/min in the dataset. This large variability in MRR values can typically be attributed to changes in processing stages or chambers, thus complicating the model learning process. Therefore, variables  $x_5$  and  $x_6$  were employed for data stratification to ensure minimal variation in MRR within each group. Prediction models were individually trained for each group to alleviate the learning complexity of the model. Table 4 illustrates the results of the data grouping. During MRR prediction, the appropriate model can be selected based on the source of process signals. Figure 8 presents the MRR distribution for each group, demonstrating a stable range.

#### Feature engineering

##### Feature extract

The aim of feature extraction is to generate features from the continuous process signal that can capture the time-dependent characteristics of the MRR. Thus, for signals such as pressure ( $X_{11} \sim X_{16}$ ), flow rate ( $X_{19} \sim X_{21}$ ) and speed ( $X_{22} \sim X_{25}$ ), elementary statistical features can be initially extracted. These

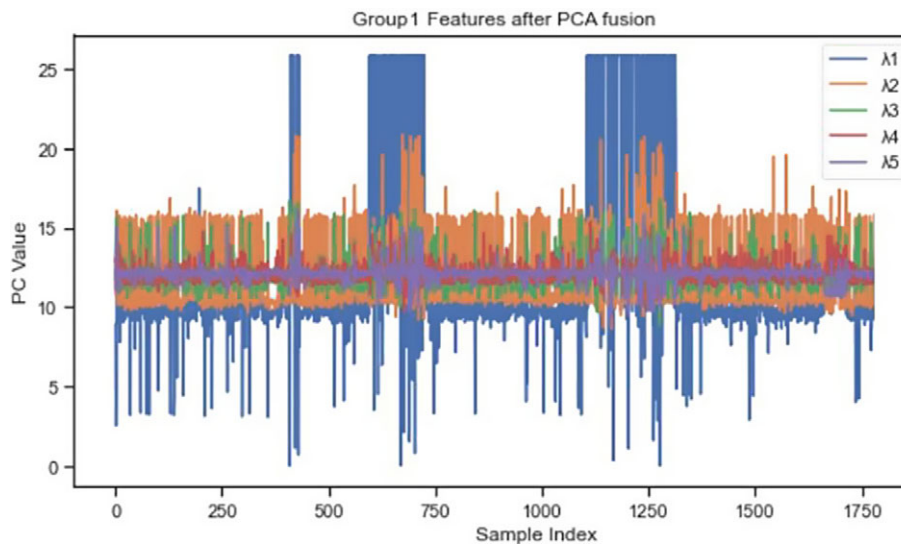
**Table 5.** High correlation feature grouping results

| Feature Group ID   | Features   |
|--------------------|--|
| G1                 | PRESSURIZED_CHAMBER_PRESSURE(mean, median, rms, std, max, pp)<br>MAIN_OUTER_AIR_BAG_PRESSURE(mean, median, rms, std, max, pp)<br>CENTER_AIR_BAG_PRESSURE(mean, median, rms, std, max, pp)<br>RIPPLE_AIR_BAG_PRESSURE(mean, median, rms, std, max, pp)<br>RETAINER_RING_PRESSURE(mean, median, rms, std, max, pp)<br>EDGE_AIR_BAG_PRESSURE(mean, median, rms, std, max, pp) |
| G2                 | SLURRY_FLOW_LINE_A(max, median, pp, std, rms, mean)<br>SLURRY_FLOW_LINE_C(median, std, rms, mean)<br>SLURRY_FLOW_LINE_B(median, rms, mean)   |
| G3                 | WAFER_ROTATION(max, rms, pp, std, mean)<br>STAGE_ROTATION(max, rms, pp, std, mean)   |
| G4                 | HEAD_ROTATION(max, rms, pp, std, mean)<br>SLURRY_FLOW_LINE_B(max, pp)  |
| G5                 | SLURRY_FLOW_LINE_C(max, pp)  |
| Ungrouped features | SLURRY_FLOW_LINE_B(std)<br>WAFER_ROTATION(median)<br><br>STAGE_ROTATION(median)<br>HEAD_ROTATION(median)   |

include Max, Mean, Median, root mean square (RMS), PEAK-TO-Peak (PP) and standard deviation (STD)—6 fundamental features in the time domain. For instance, each sample’s MRR value represents the average MRR over a period of processing time, during which multiple samplings will provide  $N$  groups of signal features. Take feature  $X_{11}$  as an example, its Max feature is the maximum value of the  $N$  groups of  $X_{11}$  features, and so forth. Subsequently, the temporal length of non-zero pressure values, indicative of normal polishing periods, can be extracted as Effective Polishing Time. Finally, given the time-series nature of the dataset, temporal neighborhood features can be employed to characterize the variation in MRR with processing time across adjacent intervals. Similarly, consumable usage neighborhood features illustrate the change in MRR with varying consumable usage. Pressure neighborhood features depict the fluctuation in MRR with altering pressure. For instance, during the process of extracting usage-based neighborhood features, the average consumables usage feature for each wafer is initially calculated during the manufacturing process. Subsequently, wafer process signals with similar usage are grouped together using the KNN clustering algorithm. Lastly, the MRR results of the wafers obtained through KNN clustering are added to the current wafer’s features as usage-based neighborhood features. As illustrated in Figures 9, 10, and 11, time, consumable usage, and pressure neighborhood features are positively correlated with MRR, respectively.

**Table 6.** Example of PCA analysis result

| Principal Component | G1 (PCV) | CCR/% | G2 (PCV) | CCR/% | G3 (PCV) | CCR/% | G4 (PCV) | CCR/% | G5 (PCV) | CCR/% |
|---------------------|----------|-------|----------|-------|----------|-------|----------|-------|----------|-------|
| $\lambda_1$         | 8.96     | 74.85 | 3.21     | 78.38 | 2.61     | 72.62 | 0.95     | 98.72 | 3.3      | 99.99 |
| $\lambda_2$         | 2.53     | 92.97 | 3.95     | 88.53 | 2.80     | 93.36 | 0.46     | 100   | 3.29     | 100   |
| $\lambda_3$         | 4.47     | 95.74 | 3.58     | 94.13 | 2.41     | 95.93 | —        | —     | —        | —     |
| $\lambda_4$         | 12.46    | 96.74 | 4.60     | 97.22 | 2.42     | 97.72 | —        | —     | —        | —     |
| $\lambda_5$         | 13.06    | 97.47 | 5.62     | 98.55 | 2.62     | 98.68 | —        | —     | —        | —     |



**Figure 13.** Group first 5 principal component values.

**Feature fusion**

Following the mentioned feature extraction procedure above, the dimensionality of the samples increased from 26 to 112. To avoid feature redundancy and the curse of dimensionality, PCCA was applied to the extracted features of slurry flow rate, rotational speed, and pressure. As depicted in Figure 12, the correlation threshold was set at 0.8, suggesting that features with correlations exceeding 0.8 could be consolidated into a single feature group. Taking the features `rms_fea_SLURRY_FLOW_LINE_A` and `mean_fea_SLURRY_FLOW_LINE_A` in Figure 12(a) as an example, they respectively represent the RMS and Mean features of the `SLURRY_FLOW_LINE_A` signal. Their PCCA value is 0.96, which exceeds the threshold of 0.8. This suggests that they can be grouped into a feature group.

Table 5 presents the results of feature grouping achieved through PCCA. Notably, Slurry B and C flow rates, which represent the flow rates of the conditioning disk and polishing pad, respectively, show weak correlations with other temporal features. Therefore, their Max and PP values form a distinct feature group. Furthermore, the STD of `SLURRY_FLOW_LINE_B` and the Median of each rotation feature display weak correlations with other features overall, and therefore, they were not included in any group.

PCA was employed to extract the principal components within groups G1 to G5. The Cumulative Contribution Rate(CCR) threshold was set at 80%. As exemplified in Table 6 with the first sample, the CCRs of the initial 2, 2, 1, and 1 principal components of groups G1 to G5 reached the predefined CCR threshold,

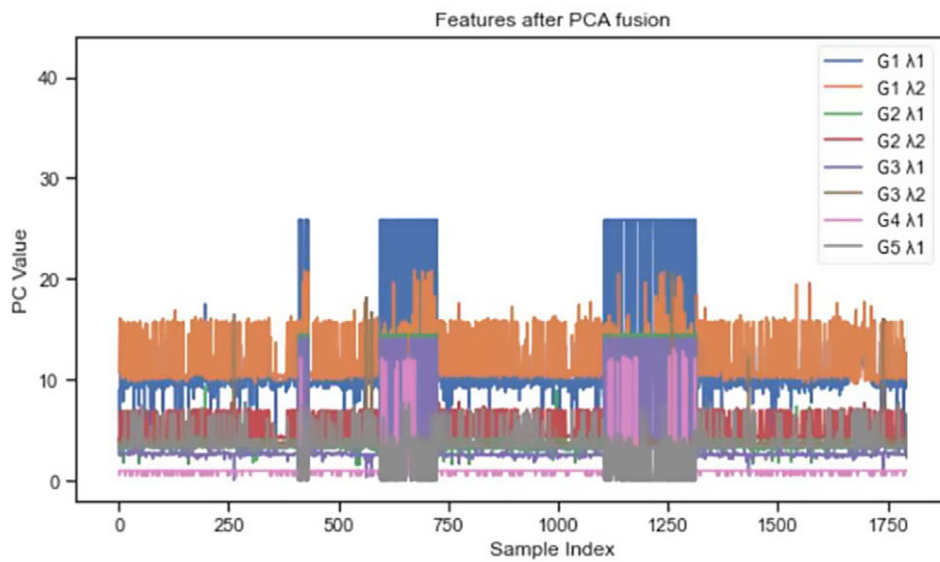


Figure 14. The final selected principal components.

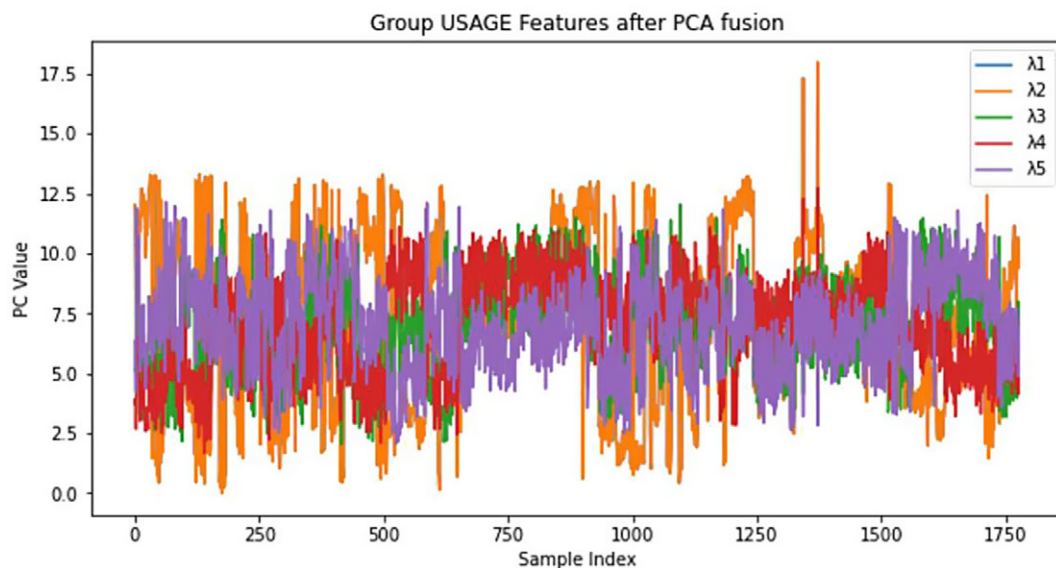


Figure 15. Visualization of the first five principal component values of consumable usage features.

**Table 7.** Example of the first five principal component values of consumable usage features

| Principal component | PCV of consumable usage features | CCR/% |
|---------------------|----------------------------------|-------|
| $\lambda_1$         | 12.08                            | 53.52 |
| $\lambda_2$         | 9.97                             | 81.20 |
| $\lambda_3$         | 4.10                             | 92.60 |
| $\lambda_4$         | 3.94                             | 96.74 |
| $\lambda_5$         | 3.84                             | 97.47 |

**Table 8.** Features after features extraction and fusion

| Types of features          | Feature dimension |
|----------------------------|-------------------|
| Effective process time     | 1                 |
| Neighboring feature        | 3                 |
| Temporal feature           | 8                 |
| Ungrouped temporal feature | 4                 |
| Consumable usage features  | 2                 |

respectively. The term “PCV” denotes the value of each principal component. Figure 13 shows the first 5 PCVs of G1,  $\lambda_1$  to  $\lambda_5$ , for each training sample. In summary, the results of principal component selection for G1 to G5 for each sample are depicted in Figure 14.

Employing a similar methodology, PCA was performed on the Consumable Usage Features ( $X_7, X_8, X_9, X_{10}, X_{17}, X_{18}$ ). As illustrated in Figure 15, the first 5 PCVs are obtained for each training sample. As exemplified in Table 7 with the first sample, the CCR of the first two principal components exceeded the CCR threshold. Consequently, the principal component features derived from the Consumable Usage Features were reduced to a two-dimensional representation.

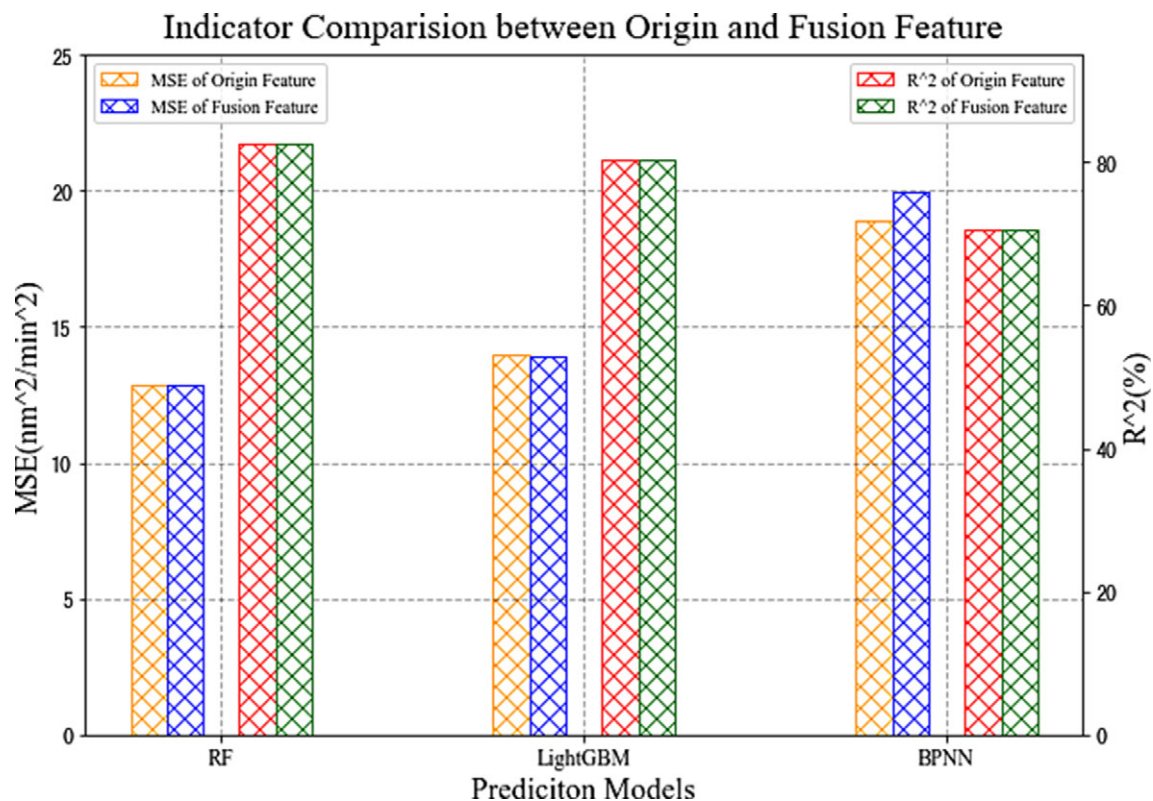
Ultimately, upon conducting feature extraction and dimensionality reduction, 18 features are acquired to serve as inputs for the prediction model, as outlined in Table 8.

### Model training and optimization

Initially, preliminary models were trained to validate the effectiveness of feature fusion. Following this, the Bayesian hyperparameter optimization algorithm (Sicard *et al.*, 2022) was employed to refine the accuracy of the preliminary models. Ultimately, by stacking these preliminary models, the MRR prediction model was constructed, thus further enhancing the prediction accuracy of MRR.

### Preliminary model prediction

On each of the three datasets partitioned in Table 3, the RF, LightGBM, and BPNN models are individually trained for MRR prediction. The accuracy of each model was denoted by the mean value of prediction precision across the three test sets. As demonstrated in Figure 16, the fused features presented lower input dimensions, at 19 compared to the original 112, while maintaining nearly the same prediction accuracy as the original features. This proves the validity of feature fusion. Low-dimensional features will

**Figure 16.** Indicator comparison between origin and fusion feature.

bring lower computational effort, thus achieving a faster inference speed without loss of accuracy.

### Hyperparameter optimization

The precision upper limit for the stacking ensemble learning model largely depends on the accuracy of the preliminary models. Hence, optimization of these preliminary models is crucial. Traditional manual optimization methods can be time-consuming, labor-intensive, and inefficient. Conversely, the Bayesian parameter optimization algorithm can efficiently utilize the information from prior function evaluations based on the Bayesian theorem, selecting the next promising sampling point as per the objective function's posterior distribution. This algorithm is highly suitable for the automated selection of model parameters. As displayed in Table 9, the parameters for each preliminary model were finalized following Bayesian parameter optimization, Value\_1, Value\_2 and Value\_3

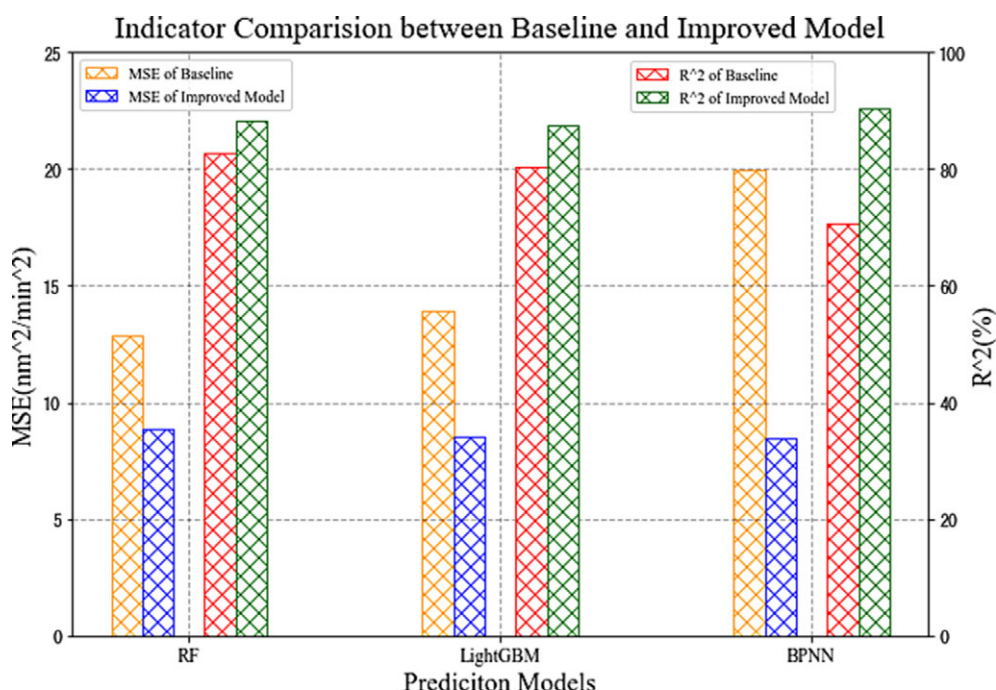
represent the model hyperparameter values corresponding to Group\_I, Group\_II and Group\_III. For the model BPNN, a three-layers neural network is selected, that is, there is only one hidden layer, and the number of its neurons is shown as 'hidden\_layer\_sizes'. Additionally, BPNN uses ReLU as the activation function and Adam as the optimizer, thereby achieving adaptive learning rate adjustment, and the initial learning rate is shown as 'learning\_rate\_init'. Similarly, the accuracy of each preliminary model was denoted by the mean values of the evaluation metrics across the three test sets. As depicted in Figure 17, the prediction accuracy of each model significantly improved post-Bayesian parameter optimization.

### Stacking model prediction

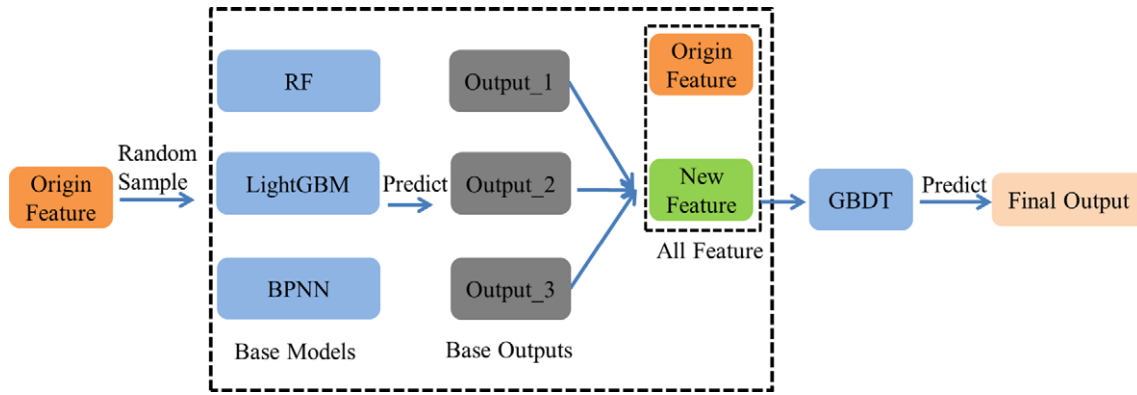
The stacking ensemble learning procedure designed is represented in Figure 18. Initially, the original fused features were input into each preliminary model to generate the prediction results. Subsequently,

**Table 9.** Model hyperparameter optimization results

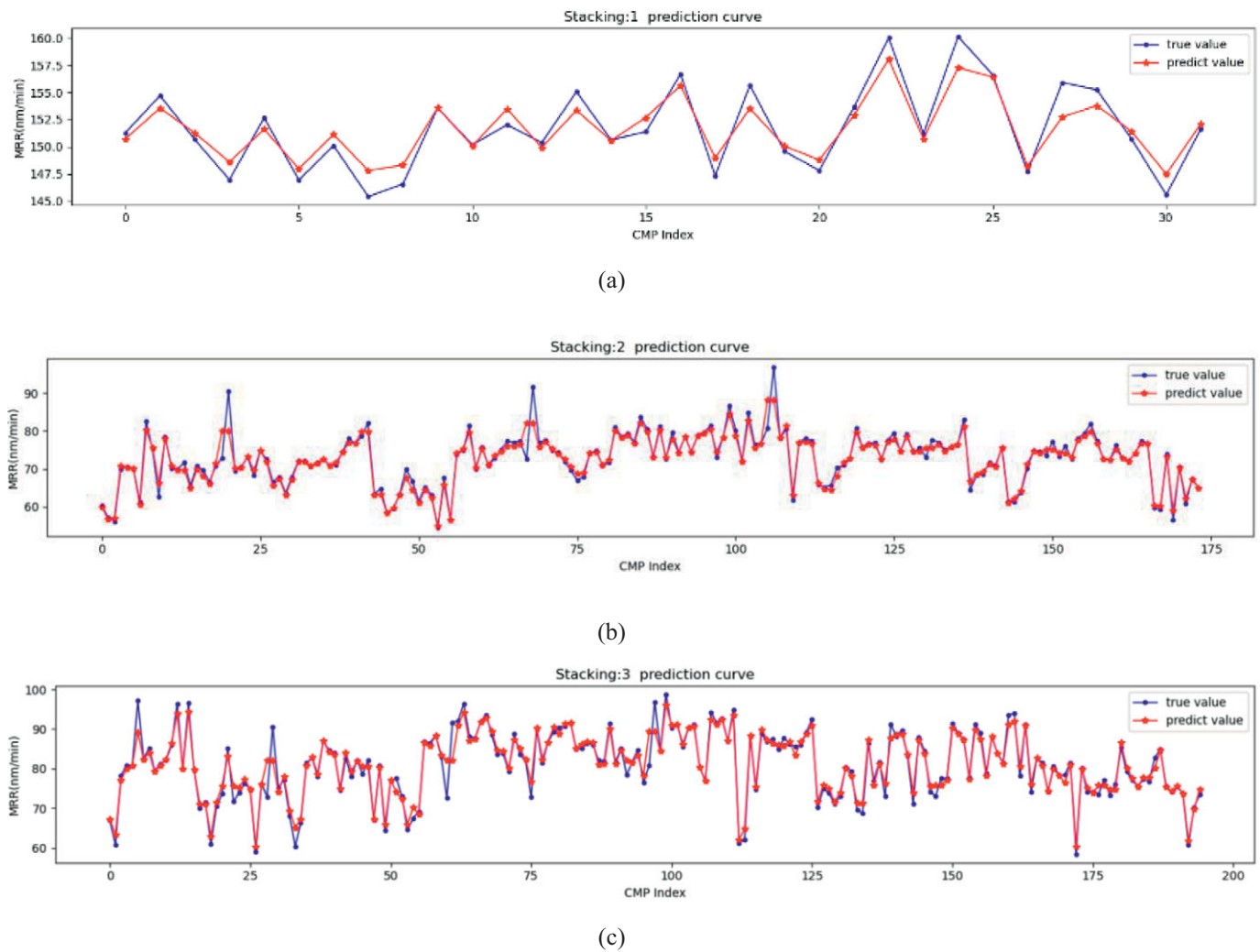
| Model    | Hyperparameters    | Range      | Value_1 | Value_2 | Value_3 | Hyperparameter description                 |
|----------|--------------------|------------|---------|---------|---------|--|
| RF       | n_estimators       | 50–500     | 105     | 405     | 280     | The number of decision trees               |
|          | max_depth          | 50–100     | 23      | 73      | 69      | The max depth of each decision tree        |
| LightGBM | n_estimators       | 50–500     | 107     | 166     | 169     | The number of decision trees               |
|          | max_depth          | 50–100     | 58      | 50      | 37      | The max depth of each decision tree        |
|          | learning_rate      | 0.001–0.05 | 0.03    | 0.04    | 0.03    | The rate of learning in each iteration     |
|          | min_child_weight   | 0.001–1    | 0.86    | 0.48    | 0.01    | The min sample weight of a child node.     |
| BPNN     | learning_rate_init | 0.001–0.05 | 0.03    | 0.03    | 0.02    | The rate of learning in each iteration     |
|          | hidden_layer_sizes | 6–14       | 6       | 7       | 9       | The number of nodes in hidden layer        |
|          | wac                | 0–0.0001   | 0.005   | 0.002   | 0.001   | The weight attenuation coefficient of Adam |



**Figure 17.** Indicator Comparison between baseline and improved model.



**Figure 18.** Training process of MRR prediction stacking model.



**Figure 19.** Predicted results on each test dataset group.

these prediction outcomes from each preliminary model were used as new features. These, along with the original fused features, were input into the secondary GBDT model. Through the GBDT’s forward process, the final MRR prediction value was produced.

Figure 19 depicts a comparison between the MRR predicted values and the actual ones from the trained Stacking MRR

prediction model over the three test sets, yielding an average MSE of 7.72. As demonstrated in Figure 20, the correlation between the model output and the actual values attained an  $R^2$  value of 95.82%. Compared with each preliminary model, the evaluation metrics show further improvement, as depicted in Figure 21.

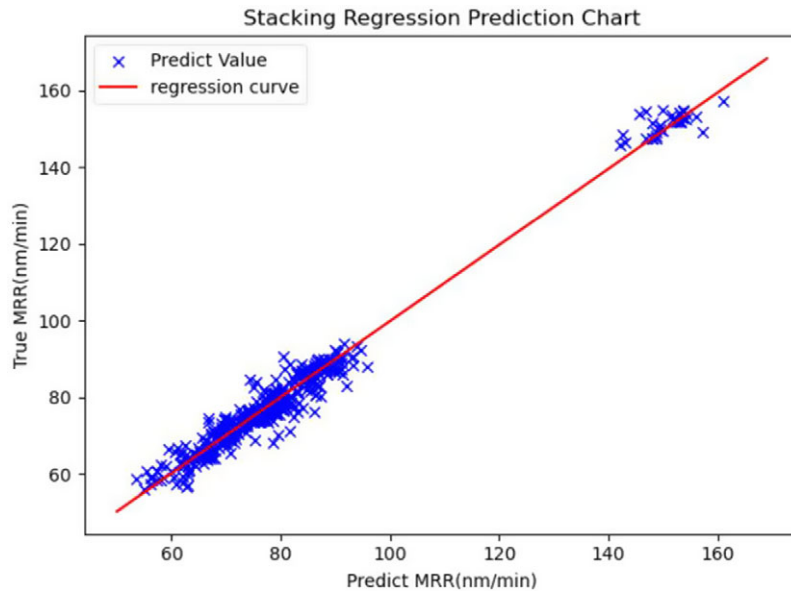


Figure 20. Correlation between prediction results and ground truth.

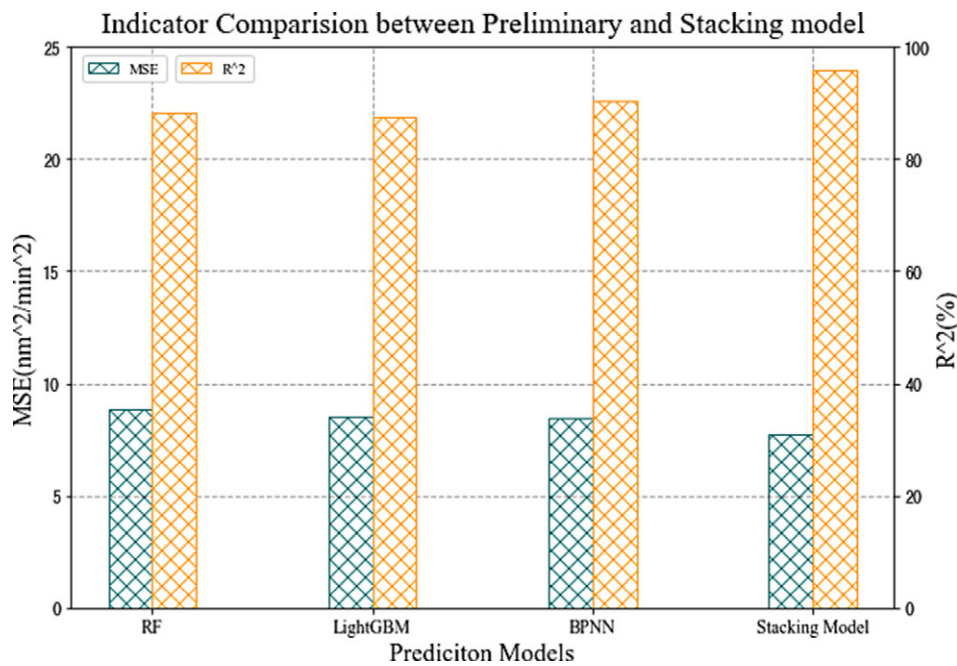


Figure 21. Comparison among stacking model and each preliminary models.

**Discussion**

As MRR prediction models of semiconductor wafer polishing process are concerned, prediction accuracy, real-time performance, and broad applicability of the application methodology are all critical. In pursuit of high prediction accuracy, this paper attempts to employ stacking models with significant structural differences, hoping that the established model can focus on data characteristics from different aspects. As depicted in Table 10, it shows the evaluation indicators of different methods on the PHM2016 dataset. Compared with simply stacking multiple tree-based models of similar structure, such as CART-Stacking and ELM-Stacking, which have MSEs of 22.88 and 21.53,

respectively, the proposed method has achieved MSE of 7.72. In comparison with the Res-CNN-based model, although the overall prediction accuracy of the stacking model is slightly lower, it has also demonstrated a certain advantage in the third group of test data.

The establishment of predictive models not only focuses on accuracy, but also emphasizes the real-time performance of prediction, which helps researchers grasp the processing status in real-time, thereby making decisions. The developed stacking model integrates tree models with efficient calculating power and a three-layer neural network model with few parameters and powerful non-linear fitting ability. Compared with models based on full CNN

**Table 10.** Comparison of models developed by different approach

| Approach  | MSE (nm/min)              | $R^2$                             |
|---|---------------------------|-----------------------------------|
| Preston model                                       | 870.25                    | –                                 |
| Physics-informed machine learning (Yu et al., 2019) | 288.08                    | –                                 |
| Luo and Dornfeld model                              | 57.76                     | –                                 |
| CART-Stacking                                       | 22.88                     | –                                 |
| ELM-Stacking  | 21.53                     | –                                 |
| Res-CNN   | 6.72 (7.48/<br>5.07/9.07) | 0.9923 (0.4448/<br>0.8803/0.8911) |
| Stacking model in this paper                        | 7.72 (7.54/7.06/8.04)     | 0.9582(0.4767/<br>0.8584/0.8990)  |

structure, it achieves a balance between precision and speed. Furthermore, the broad applicability determines whether the method can be conveniently and effectively applied to other scenarios and is key to industrialization. As described above, the application of the Res-CNN-based model is, to a certain extent, limited by the dimensions of the input features. The developed stacking model can accept inputs of any dimension and can be more conveniently migrated to other scenarios.

Despite that, there is still room for improvement in terms of speed and accuracy for the developed MRR prediction model. Hence, as machine learning techniques evolve, more efficient and accurate model structures will emerge, such as the recently popular Transformer series models (Vaswani et al., 2017; Devlin et al., 2019; Dosovitskiy et al., 2020). Exploring how to apply these emerging models with novel structural forms in model fusion to push the boundaries of MRR prediction accuracy may be a worthwhile direction for future research.

## Conclusion

In this paper, a CMP MRR prediction model for semiconductor wafers that integrates multiple preliminary models with significant structural differences is developed, utilizing the stacking ensemble learning method. The experiments were conducted on the PHM2016 dataset, which involved the analysis and preprocessing of raw data, followed by feature extraction and fusion. The resulting fused features served as input to train both preliminary and stacking models, and their effectiveness was validated using a test set. The main conclusions from this study are as follows:

- (1) A feature extraction and fusion pipeline was created for the semiconductor wafer CMP process signals, relying on the PCCA and PCA. This method effectively reduced the extracted 112-dimensional features to 18 dimensions, without compromising the prediction accuracy of the model. It demonstrated potential in reducing the computational load of the model and enhancing the real-time performance of MRR prediction.
- (2) A CMP MRR prediction model for semiconductor wafers was developed using the PHM2016 dataset. Compared with the preliminary models, the final prediction model showed further improvement in accuracy, reducing the MSE to 7.72 and

raising the  $R^2$  value to 95.82%. These results validate the efficacy of the data-driven method in constructing the MRR prediction model.

- (3) For the first time, an ensemble learning method has been employed to integrate multiple preliminary models with significant structural and principle differences into the development of a data-driven CMP MRR prediction model for semiconductor wafers. Compared to existing studies that incorporate preliminary models with similar structures or principles, our approach achieved higher prediction accuracy. This sets the stage for merging a wider array of efficient and diverse preliminary models in the future, aiming to push the boundaries of MRR prediction precision.

**Data availability.** The data that support the findings of this study are openly available at <https://www.phmsociety.org/sites/phmsociety.org/files/2016%20PHM%20DATA%20CHALLENGE%20CMP%20DATA%20SET.zip>

**Funding statement.** This research was supported by the financial support from the National Natural Science Foundation of China (U20A20293 and 52175441), and the Natural Science Foundation of Zhejiang Province (LD22E050010).

**Competing interest.** The author(s) declare none.

## References

- Batista GEAPA, Prati RC and Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 20–29. <https://doi.org/10.1145/1007730.1007735>.
- Breiman L (2001) Random Forests. *Machine Learning* 5–32. <https://doi.org/10.1023/a:1010933404324>.
- Devlin J, Chang M-W, Lee K and Toutanova K (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North. Presented at the Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota*. <https://doi.org/10.18653/v1/n19-1423>
- Di Y, Jia X and Lee J (2021) Enhanced virtual metrology on chemical mechanical planarization process using an integrated model and data-driven approach. *International Journal of Prognostics and Health Management* 8(2). <https://doi.org/10.36001/ijphm.2017.v8i2.2641>.
- Dosovitskiy A, Beyer, L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, ... Housby N (2020) *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. *arXiv: Computer Vision and Pattern Recognition*.
- Evans CJ, Paul E, Dornfeld D, Lucca DA, Byrne G, Tricard M and Mullany BA (2003) Material removal mechanisms in lapping and polishing. *CIRP Annals* 52(2), 611–633. [https://doi.org/10.1016/s0007-8506\(07\)60207-8](https://doi.org/10.1016/s0007-8506(07)60207-8).
- Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. <https://doi.org/10.1214/aos/1013203451>
- Hanin B and Rolnick D (2019) Deep ReLU networks have surprisingly few activation patterns. *Neural Information Processing Systems, Neural Information Processing Systems*.
- Jia X, Huang B, Feng J, Cai H and Lee J (2021) A review of PHM data competitions from 2008 to 2017: Methodologies and analytics. *Annual Conference of the PHM Society* 10(1). <https://doi.org/10.36001/phmconf.2018.v10i1.462>.
- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, ... Liu T-Y (2017) LightGBM: a highly efficient gradient boosting decision tree. *Neural Information Processing Systems*.
- Köksoy O (2006) Multiresponse robust design: Mean square error (MSE) criterion. *Applied Mathematics and Computation* 175(2), 1716–1729. <https://doi.org/10.1016/j.amc.2005.09.016>.
- Lee H (2019) Semi-empirical material removal model with modified real contact area for CMP. *International Journal of Precision Engineering and Manufacturing* 20(8), 1325–1332. <https://doi.org/10.1007/s12541-019-00161-6>.
- Li X, Wang C, Zhang L, Mo X, Zhao D and Li C (2018) Assessment of physics-based and data-driven models for material removal rate prediction in



- chemical mechanical polishing. In *International Conference on Electrical Engineering and Automation (ICEEA 2018)*. Chengdu, China. <https://doi.org/10.2991/iceea-18.2018.26>.
- Li Z, Wu D and 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). <https://doi.org/10.1115/1.4042051>.
- Malik M, Nehra AK and Saini BK** (2021) A study on factors affecting job satisfaction of working women with Karl Pearson's chi-square test. *Research Journal of Humanities and Social Sciences* **12**(2), 5–11.
- Pearson P and Karl K** (2010) *LIII. On lines and planes of closest fit to systems of points in space*. Philosophical Magazine Series 1, Philosophical Magazine Series 1.
- Ruan B** (2021) Prediction of stock market by BP neural network model. *Journal of Physics: Conference Series*, 042232. <https://doi.org/10.1088/1742-6596/1744/4/042232>
- Rumelhart DE, Hinton GE and Williams RJ** (1986) Learning representations by back-propagating errors. *Nature* 533–536. <https://doi.org/10.1038/323533a0>
- Sicard D, Briois P, Billard A, Thevenot J, Boichut E, Chapellier J and Bernard F** (2022) Deep Learning and Bayesian Hyperparameter Optimization: A Data-Driven Approach for Diamond Grit Segmentation toward Grinding Wheel Characterization. *Applied Sciences* **12**(24), 12606. <https://doi.org/10.3390/app122412606>.
- Sun Y, Li H, Zhao X, Fei J, Liu X and Niu Y** (2022) A Novel Denoise Method of Acoustic Signal from Train Bearings Based on Resampling Technique and Improved Crazy Climber Algorithm. *Shock and Vibration* **2022**, 1–11. <https://doi.org/10.1155/2022/8303722>.
- Tang R, Tao Y, Li J, Chen Z, Deng X and Li H** (2022) The Short-time Prediction of the Energetic Electron Flux in the Planetary Radiation Belt Based on Stacking Ensemble-Learning Algorithm. *Space Weather*. <https://doi.org/10.1029/2021sw002969>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez Aidan N and Polosukhin I** (2017) Attention is All you Need. *Neural Information Processing Systems*.
- Wang P, Gao RX and Yan R** (2017) A deep learning-based approach to material removal rate prediction in polishing. *CIRP Annals* **66**(1), 429–432. <https://doi.org/10.1016/j.cirp.2017.04.013>.
- Wolpert David H** (1992) Stacked generalization. *Neural Networks* (2). [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
- Xu Q, Chen L, Cao H and Liu J** (2021) A neural network-based approach to material removal rate prediction for copper chemical mechanical planarization. *ECS Journal of Solid State Science and Technology* **10**(5), 054003. <https://doi.org/10.1149/2162-8777/abfc20>.
- Xu Q, Chen L, Liu J and Cao H** (2020) A wafer-scale material removal rate model for chemical mechanical planarization. *ECS Journal of Solid State Science and Technology* **9**(7), 074002. <https://doi.org/10.1149/2162-8777/abadea>.
- Yu J-h, Lin Y-j, Zhang B, Qu Y-x, Wang B-q, Li Z-r, Xia Y-c and Chen L** (2022) Prediction method of premixed flammable gas explosion experimental result based on Adam-BP. *Journal of Dalian Maritime University* (**02**), 110–117. <https://doi.org/10.16411/j.cnki.issn1006-7736.2022.02.013>
- Yu T, Li Z and Wu D** (2019) Predictive modeling of material removal rate in chemical mechanical planarization with physics-informed machine learning. *Wear* **426–427**, 1430–1438. <https://doi.org/10.1016/j.wear.2019.02.012>
- Zhang J, Jiang Y, Luo H and Yin S** (2021) Prediction of material removal rate in chemical mechanical polishing via residual convolutional neural network. *Control Engineering Practice* 104673. <https://doi.org/10.1016/j.coneng-prac.2020.104673>.
- Zhang Z, Liu J, Hu W, Zhang L, Xie W and Liao L** (2021) Chemical mechanical polishing for sapphire wafers using a developed slurry. *Journal of Manufacturing Processes* **62**, 762–771. <https://doi.org/10.1016/j.jmapro.2021.01.004>.
- Zhao Y and Chang L** (2002) A micro-contact and wear model for chemical-mechanical polishing of silicon wafers. *Wear* **252**(3–4), 220–226. [https://doi.org/10.1016/S0043-1648\(01\)00871-7](https://doi.org/10.1016/S0043-1648(01)00871-7).
- Zhou H, Wang X and Zhu R** (2022) Feature selection based on mutual information with correlation coefficient. *Applied Intelligence* **52**(5), 5457–5474. <https://doi.org/10.1007/s10489-021-02524-x>.
- Zounemat-Kermani M, Stephan D, Barjenbruch M and Hinkelmann R** (2020) Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models. *Advanced Engineering Informatics* **43**, 101030. <https://doi.org/10.1016/j.aei.2019.101030>.

**Author biographies.** **Zhi-Long Song, Ph.D.** is a candidate at the School of Mechanical Engineering, Zhejiang University of Technology. His research focuses on intelligent manufacturing, precision, and ultra-precision machining technology.

**Wen-Hong Zhao** is a professor at the School of Mechanical Engineering, Zhejiang University of Technology. His main research focus is on ultra-precision machining and control.

**Xiao Zhang** is a master's degree student at the School of Mechanical Engineering, Zhejiang University of Technology. His main research focus is on intelligent manufacturing, ultra-precision machining and control.

**Ming-Feng Ke** is a Ph.D. candidate at the School of Mechanical Engineering, Zhejiang University of Technology. His research focuses on intelligent manufacturing, precision, and ultra-precision machining technology.

**Wei Fang** is a master's degree candidate at the School of Mechanical Engineering, Zhejiang University of Technology. Her research focuses on precision and ultra-precision abrasive machining technology.

**Bing-Hai Lyu, Ph.D.** is a professor at the School of Mechanical Engineering, Zhejiang University of Technology. His main research focus is on precision and ultra-precision abrasive machining technology.