Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-025-02753-8

Process data-driven machine learning for non-uniformity prediction and virtual metrology in chemical mechanical planarization

Morten Breidung[1,2] · Tom Rothe[1,3] · Andre Lauff[2] · Peter Thieme[2] · Jan Langer[3] · Manuel Günther[2] · Harald Kuhn[1,3]

Received: 12 September 2025 / Accepted: 22 November 2025
© The Author(s) 2025

Abstract

Chemical mechanical planarization is an integral part of semiconductor industry. It provides wafersurface smoothing at the nanometer scale and is highly monitored. In a typical industrial fabrication facility, millions of data points are generated annually for process control, fault detection and classification purposes. By leveraging this sensor and process data for the training of machine learning models, a foundation for virtual metrology can be established. Utilizing real-world data from a high-mix, high-volume fab, we developed robust and application-oriented machine learning models capable of predicting non-uniformity and spatially resolved material removal rates on wafer surfaces. Experimental results demonstrate the accuracy of these predictions, with a high degree of precision in estimating spatially resolved material removal rates. Furthermore, the analysis of the feature importance via Shapley Additive Explanations of the models reveals that polishing time and carrier rotation are among the most critical factors influencing MRR variability.

Keywords Virtual metrology · Chemical mechanical planarization · Semiconductor Manufacturing · Machine Learning

Introduction

Semiconductor devices are the backbone of modern electronics, enabling computing, communication, and automation advancements. As device dimensions shrink and performance demands increase, the precision of manufacturing processes becomes ever more critical. Chemical mechanical planarization (CMP) is a key step in semiconductor device fabrication, ensuring ultra-flat surfaces necessary for subsequent lithography and thin-film deposition steps.

CMP operates through a synergistic combination of chemical etching and mechanical abrasion, as shown in Fig. 1. The process involves pressing a rotating wafer face down onto a polishing pad. At the same time, an abrasive slurry, a colloidal suspension of nanoparticles (e.g., silica or alumina), flows between the wafer and pad. The pad is mounted onto the platen, which rotates in the same direction as the carrier (polish head). The slurry's chemical components (e.g., oxidizers) weaken the bonds of the wafer's atomic surface layer, while mechanical friction from the pad and abrasive particles removes material. The retaining ring holds the wafer in place during the polishing process. Residual by-products and particulate waste are removed by flushing the apparatus. The conditioner regenerates the pad

✉ Morten Breidung
morten.breidung@s2024.tu-chemnitz.de

Tom Rothe
tom.rothe@etit.tu-chemnitz.de

Andre Lauff
andre.lauff@infineon.com

Peter Thieme
peter.thieme@infineon.com

Jan Langer
jan.langer@enas.fraunhofer.de

Manuel Günther
manuelfrederik.guenther@infineon.com

Harald Kuhn
harald.kuhn@enas.fraunhofer.de

1 Faculty of Electrical Engineering and Information Technology, TU Chemnitz, Reichenhainer Straße 70, 09111 Chemnitz, Saxony, Germany

2 Process Development, Infineon Technologies Dresden AG & Co. KG, Koenigsbruecker Str. 180, 010199 Dresden, Saxony, Germany

3 Fraunhofer Institute for Electronic Nano Systems (ENAS), Technologie-Campus 3, 09126 Chemnitz, Saxony, Germany

Published online: 08 December 2025

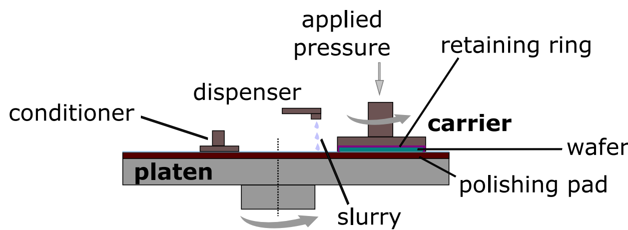

Fig. 1 A schematic overview of the CMP process

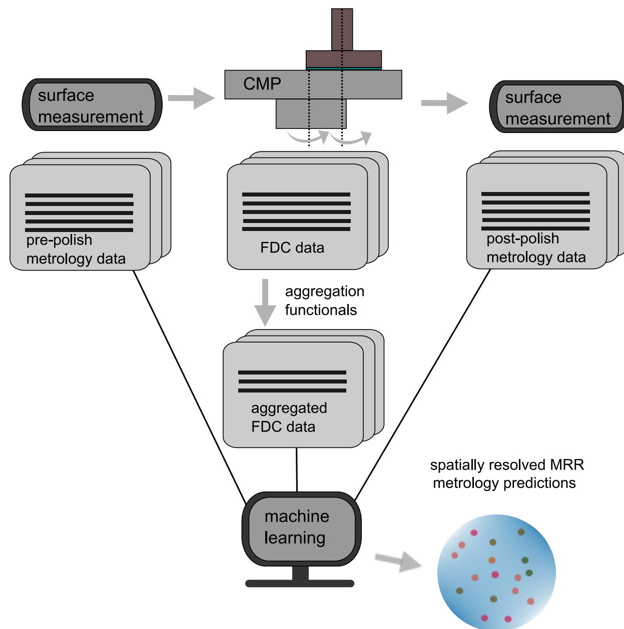

Fig. 2 Illustrates a schematic approach of the experiments section. A selection of the wafers' surface measurements serves as a proxy for the target parameters. The sensor time series are aggregated to derive features that are used for ML training. The metrology target prediction is our approach to VM

after polishing. Consumables, namely the slurry, the polishing pad, the conditioner, and the retaining ring, are typically replaced or replenished as they degrade during operation.

Maintaining consistent material removal rates (MRR) across the wafer surface and achieving low within-wafer non-uniformity (WIWNU) is a critical challenge in CMP. Non-uniformity can cause significant issues such as excessive topography, interlayer dielectric erosion, and dishing, directly affecting yield, reliability, and downstream lithography accuracy (Shauly and Rosenthal (2020)).

To control material removal after the CMP process, time-consuming and costly metrological measurements are conducted both pre- and post-polishing. However, the current methodology relies on sampling. To reduce costs and time, virtual metrology (VM) is considered a key solution to surpass the physical metrology method. VM predicts important metrics in semiconductor manufacturing (e.g., film thickness, uniformity, defects) without physical measurements. While existing VM models estimate averaged MRRs, they lack the spatial resolution for uniformity control. This limitation is particularly prominent for multi-zone pressure heads, where the interfacial pressure between wafer and pad can be spatially fine-tuned. To address this critical gap, this paper aims to predict spatially resolved MRRs and the WIWNU for VM purposes.

Figure 2 illustrates the schematic approach of this study. We utilize real-world CMP process data provided by Infineon. The targets for the VM are obtained via spectroscopic reflectometry measurements before and after polishing on 17 chosen measurement sites on the wafer surface. We aggregate sensor time series data to create sensor-derived features. Three regression models are employed: linear regression (LR), random forest regression (RFR), and extreme gradient boosting regression (XGBoost). We use a comprehensive dataset of more than 13,000 unique wafer runs to train the VM models and we evaluate the model's performances and investigate the features' importance to enable explainability.

A WIWNU prediction model serves as a general predictor of wafer non-uniformity, whereas the spatially resolved MRR prediction models provide predictions for defined regions on the wafer surface. Our approach aims to reduce the dependency on direct wafer measurements while improving process control in high-mix, high-volume fabs, by providing the control system with additional 'virtual' measurements.

A more in-depth investigation of the current challenges of VM and the literature is presented in Sect. 2. The target selection for VM, the aggregation of sensor time series data, and the addition of features to enhance model performance are described in Sect. 3. The Section also outlines the theoretical background of the applied regression models. Section 4 describes the experimental implementation of the target and feature design and the code and models deployed. In Sect. 5, we evaluate the performance of the models and analyze the feature importance. Section 6 summarizes the key findings and outlines directions for future work.

Related work

Establishing a uniform surface is commonly managed by adjusting the pressure distribution through multi-zone polishing heads, which apply localized pressure in concentric zones (Shiul et al. (2004)). Traditional methods for determining optimal pressure distributions are based on Design of Experiments (DoE). However, as the number of pressure zones increases in state-of-the-art polishing heads (Wang et al. (2013)), DoE approaches become impractical due to the extensive wafer usage required. Those radius-specific

pressure heads provide fine-tuning options to ensure planarization at the nanometer scale. However, they require continuous tuning throughout the manufacturing process. To mitigate wafer consumption, research has focused on developing models that interpolate DoE results using high-fidelity simulations, machine learning (ML) techniques (Yi et al. (2003); Cho et al. (2025)) or hybrid modeling strategies (Rothe et al. (2025)).

Although these model-based approaches effectively reduce initial wafer consumption, ongoing control of wafer uniformity in production still relies on Run-to-Run (R2R) adjustments (Shiul et al. (2004)), allowing pressure tuning for individual wafers. However, the high costs and temporal offsets associated with metrology limit the practicality of extensive physical measurements. Typically, only a small subset of wafers per batch is measured despite the availability of Fault Detection and Classification (FDC) sensor data for every wafer.

To overcome limited metrological coverage VM has emerged as an effective solution. It leverages sensor data to predict measurement outcomes, enhancing R2R control and enabling early detection of process drifts. ML is increasingly preferred in VM applications due to its ability to model the complex and nonlinear relationships inherent in CMP processes, significantly outperforming traditional statistical methods (Djedidi et al. (2022)). Reflecting this trend, a recent literature review from Winkler et al. (2025) covering publications up to 2024 identified 81 ML-focused CMP studies, with 45 specifically addressing VM.

While diverse in scope, these studies tend to follow a typical methodological pattern (cf. Winkler et al. (2025)). It is common to train separate models for different groups (e.g., chambers or products), though this often limits generalizability. Outliers and samples with missing targets are typically discarded, and continuous sensor signals are reduced to statistical descriptors (e.g., mean, standard deviation).

Categorical variables are usually one-hot encoded, and normalization is standard. Feature engineering strategies frequently include temporal neighbors (e.g., preceding wafer runs), feature-space neighbors (similar process conditions), and physically derived features Xia et al. (2021), such as those based on Preston's equation.

Dimensionality reduction and feature selection methods are usually applied to manage high-dimensional inputs. Many algorithms have been explored, from classical models (e.g., linear regression, random forests) to deep learning Lee and Kim (2018), Bayesian approaches, and ensemble methods Wei and Wu (2022). Recent work shows ensemble models consistently outperform individual learners (Li et al. (2019)), with XGBoost emerging as a top performer (Deivendran et al. (2025)).

To capture temporal patterns of sensor reading time-series, sequence-aware features or networks can be applied. A common middle ground is a sliding-window preprocessing, for example, segment each run into overlapping time windows and compute the same statistics on each window. This retains some chronological context and often yields modest accuracy gains. At the far end are deep learning models (1-D CNNs, LSTMs, Transformers) that directly ingest raw time-series and learn feature embeddings. These can capture complex dynamics and long-range dependencies, potentially improving predictions. For example, on the Prognostics and Health Management Society (2016) PHM 2016 CMP dataset, the baseline XGBoost (using run-level aggregates) by Liu et al. (2022) achieved an MAE of ≈ 1.94. In contrast, a 1-D CNN (trained end-to-end on the raw sensor sequences) yielded the best accuracy (MAE ≈ 1.87). Although the authors did not specify the computational costs for this model, we can assume that it introduces far greater complexity: e.g., many more parameters, longer training, as well as inference time, and the need for specialized hardware or software. Farahani et al. (2025) showed that transformer architectures can outperform simpler models in some manufacturing time-series problems, but often with only marginal error reduction. In practice, deploying a CNN or Transformer in a fab requires significant engineering effort (streaming data interfaces, real-time inference on device, etc.), whereas the XGBoost on sensor reading aggregates fits readily into existing infrastructures. The trade-off between model complexity and predictive performance is evident: relying solely on statistical aggregates yields a parsimonious model with straightforward deployment, whereas sliding-window or deep learning approaches can extract marginally greater predictive power, albeit at the cost of substantially increased computational and integration overhead.

Despite these advances, critical limitations remain unaddressed, particularly in the following areas:

Spatially resolved predictions: Most VM studies focus on spatially averaged targets, such as mean MRR or post-polish thickness, while ignoring spatial distributions across the wafer. However, spatially resolved outcomes are essential for monitoring and controlling within-wafer uniformity. To date, no ML-based model has been developed to predict full-surface MRR or thickness maps from FDC data, leaving a critical gap for enabling R2R control of uniformity. This represents the central challenge addressed in this work. Spatially resolved predictions for 17 distinct points or a full wafer map enable enhanced R2R control strategies that are not feasible with mean MRR predictions alone. Specifically, these predictions allow for targeted adjustments in multi-zone pressure heads, addressing localized variations in material removal and improving process uniformity. For

instance, if the prediction model indicates higher removal near the wafer edge, the R2R controller can proactively reduce edge-zone pressure while maintaining center pressure, thereby improving uniformity. Similarly, systematic spatial patterns detected across wafers can be used to update zone-specific setpoints, leading to a more targeted compensation strategy than with global corrections alone.

Real-world data utilization: Real-world data usage is complicated by the highly specialized and individualized nature of wafer polishing, particularly in high-mix manufacturing settings, where multiple technologies are produced concurrently.

Many existing studies rely on synthetic or legacy datasets, such as the widely used PHM dataset from the Prognostics and Health Management Society (2016), which includes only 2, 398 wafer runs and 25 features. While valuable for benchmarking, these datasets do not reflect real production environments' scale, variability, and feature richness. Some proprietary datasets from STMicroelectronics (Jebri et al. (2016); Djedidi et al. (2022)) include up to 149 FDC and historical features, such as post-etching depth or post-CVD thickness, but are limited in size, typically covering just 545 to 1, 500 wafers. Only Tsuda et al. (2015) report using data from an entire year of production without disclosing the actual number of wafers, making it difficult to evaluate the scale and representativeness of their study. Their approach focuses on predicting a single global removal rate per wafer and relies on multiple linear regression, which cannot fully capture nonlinear effects. In contrast, this work uses 13, 675 wafer runs and 163 derived features to predict the spatially resolved MRR distribution across the wafer with modern machine learning algorithms.

Model explainability: Despite its importance for trust, validation, and model debugging, explainability remains largely overlooked in CMP-related ML research. The mostly interconnected aspects of the CMP operation make it complex to model the whole process. Physics-informed models are increasingly in the focus of CMP research. To predict MRR, hybrid models leverage domain knowledge, such as the relationship between slurry pH, particle size, and scratch formation (Li et al. (2019); Rahman et al. (2024)). By leveraging sensor data, ML algorithms can uncover process and tool dependencies that are difficult to capture using conventional methods by investigating feature importance. This can inform the physics-based approach. Only a few studies make explainability a priority. A notable exception is the work by Zuo et al. (2024), who applied SHAP (SHapley Additive exPlanations) to an XGBoost model to quantify feature importance and provide insight into the model's decision-making process. In this work, we follow a similar approach and use SHAP to analyze feature importance in our XGBoost-based VM model.

In summary, we extend existing CMP VM approaches by (i) estimating MRRs with spatial resolution across sites, (ii) demonstrating generalization on heterogeneous real-world operational data, and (iii) integrating SHAP-driven explanations to enable interpretable process optimization.

Methodology

This section presents the theoretical foundations of the employed feature selection, regression modeling, and SHAP analysis.

Feature selection using mutual information

The feature selection technique combines the mutual information (MI) between the features X_i and the targets Y_j with the distance entropy estimates of the k-nearest neighbors.

Mutual Information (MI) quantifies the statistical dependence between two random variables. (Kraskov et al. (2004)) MI helps identify the most informative features by measuring how much knowledge of a feature reduces uncertainty about Y_j. It is defined as

$$I(X_i; Y_i) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right), \tag{1}$$

where $p(x, y)$ is the joint probability distribution, and $p(x)$, $p(y)$ are the marginal distributions.

After computing $I(X_i; Y_i)$ for each feature X_i using k-nearest neighbors, the features are ranked by their MI scores. Then, the top-k features with the highest MI are selected.

Theory of regression models

Random forest regression

A random forest regression is an ensemble learning method that builds multiple decision trees during training and aggregates their outputs for the final prediction (Breiman (2001)). Constructing a decision tree involves recursively partitioning the data by selecting a random subset of variables to split a parent node into two child nodes. The splitting criterion for each parent node can be formulated as an optimization problem that seeks to minimize the sum of squared errors within the two resulting regions

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

where R_1 and R_2 denote the two regions resulting from the splitting process, defined as $R_1 = x|x_j \leq c$ and $R_2 = x|x_j \geq c$, respectively, with the splitting threshold c. Here, x_j represents the j-th splitting variable, c is the cutting point, and m_1 and m_2 are the means of the values y_i within the regions $R1$ and $R2$ (Breiman (2001)).

XGBoost regression

XGBoost, introduced in 2016 by Chen and Guestrin (2016), is an ensemble learning method that combines multiple weak learners, typically decision trees, to improve prediction accuracy. Trees are constructed sequentially, with each new tree aiming to correct the errors of the previous ones. During training, each tree is assigned a weight based on its contribution to reducing the overall prediction error.

The model optimizes the following objective function

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

where y_i is the true target value, $\hat{y}_i$ is the predicted value, L is a differentiable loss function (e.g., squared error or cross-entropy), and $\Omega(f_k)$ is a regularization term that penalizes tree complexity to prevent overfitting. Specifically, $\Omega(f_k)$ is defined as

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2, \tag{4}$$

where T is the number of leaves in the tree, ω_j is the weight of leaf j, and γ and λ are regularization parameters controlling the penalties for tree complexity and leaf weights, respectively. By including $\Omega(f_k)$, XGBoost balances the fit to the training data with model simplicity, improving generalization performance.

Performance metrics

The coefficient of determination (R^2) quantifies the proportion of variance in the target variable that the model explains. It is computed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{5}$$

where the numerator is the residual sum of squares (SSE) and the denominator is the total sum of squares (SST), y_i is the observed value of the dependent variable, $\hat{y}_i$ is the predicted value of the dependent variable, and n is the number of observations.

The mean absolute percentage error (MAPE) measures the average relative error between predicted values $\hat{y}_i$ and ground truth y_i, expressed as a percentage. MAPE is undefined when $y_i = 0$ and can be heavily biased when y_i is close to zero, which may distort the model evaluation. It is defined as

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100. \tag{6}$$

In addition to these relative measures, the absolute error (AE) provides an intuitive measure of the magnitude of the prediction error in the same units as the target variable. It is defined as

$$\text{AE}_i = |y_i - \hat{y}_i|. \tag{7}$$

While the mean or median of the absolute error reflects the central tendency of the prediction deviations, these statistics can obscure extreme deviations. To quantify the model's performance under worst-case or near-worst-case scenarios, the 95th and 99th percentile absolute errors (95th pct. and 99th pct. AE) are used. These represent the error magnitudes below which 95% and 99% of all predictions fall. These high-percentile metrics are particularly valuable for assessing model robustness and reliability in a manufacturing environment. In contrast to the mean-based metrics, which summarize overall accuracy, percentile-based metrics capture the tail behavior of the error distribution and thus complement the interpretation of model performance in critical process windows.

Nested cross-validation

Nested cross-validation (NCV) is a well-established approach for hyperparameter optimization. NCV consists of two loops: an outer loop and an inner loop, as shown in Fig. 3. The outer loop partitions the labeled dataset into 10 folds. In each iteration, one fold is held out for validation, while the remaining nine folds are passed to the inner loop. The inner loop performs a three-fold cross-validation on these nine folds, where hyperparameters are optimized via randomly sampled grid search. The best hyperparameters are then used to retrain the model on all inner loop data, and the performance is evaluated on the outer validation fold. This process is repeated across all outer folds, yielding 10 independently optimized models. The final performance (R^2) is reported as the mean of the 10 outer validation scores. Since

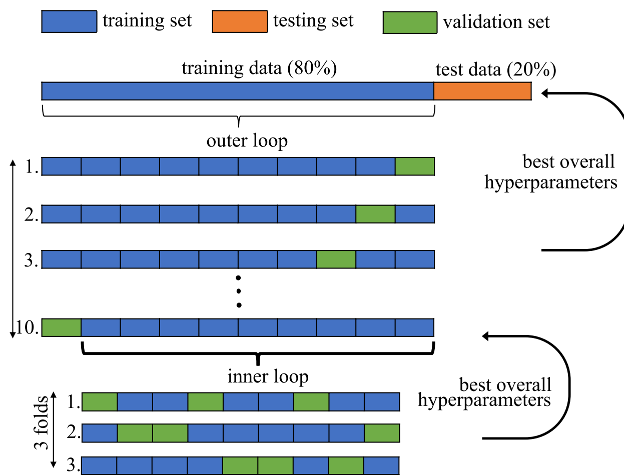

Fig. 3 Illustrates the data splitting for the NCV. The dataset undergoes an initial time-based split (80:20) to separate test data for the feature analysis and test the models' ability to generalize over time

no data is reused improperly, NCV is less prone to overfitting in limited-data scenarios than standard cross-validation.

SHAP

SHapley Additive exPlanations was introduced in 2017 by Lundberg and Lee (2017). Given a machine learning model *f* and a set of features X_i, the SHAP value ϕ_i for feature *i* is computed as

$$\phi_i = \sum_{S \subseteq X_i \setminus \{i\}} \frac{|S|!\,(|X_i| - |S| - 1)!}{|X_i|!} \left(f(S \cup \{i\}) - f(S)\right). \quad (8)$$

S is a subset of features excluding *i*, *f*(*S*) is the model's prediction using only the features in *S*. The weighting factor $\frac{|S|!\,(|F|-|S|-1)!}{|F|!}$ ensures fair attribution. Following the game-theoretic approach proposed by Shapley (1997), each feature *X* is viewed as a "player" contributing to the "payout" (the prediction), where the goal is to assign credit to each feature for the difference between the prediction and the baseline (average prediction).

Tree SHAP is a variant of SHAP optimized for tree-based machine learning models. Using the recursive structure of decision trees reduces the computation time through three mechanisms. First, each tree is traversed recursively. Second, the algorithm tracks the proportion of feature subsets in which a given feature affects the prediction. Third, contributions are aggregated across all trees in the ensemble. This approach enables fast and interpretable feature attribution (Lundberg et al. (2018)).

Table 1 Aggregation functionals applied for sensor time series aggregation

Type	Numpy Implementation
Area under the curve	`trapz()`
Maximum	`max()`
Minimum	`min()`
Mean	`mean()`
Peak to peak	`ptp()`
Mean of the derivative	`diff().mean()`
Maximum of the derivative	`diff().max()`
Minimum of the derivative	`diff().min()`

Experiments

The primary objective of this study is to predict spatially resolved MRRs and WIWNU from heterogeneous process data, enabling VM. By adding sensor data, process parameters, and consumable states, the models aim to reduce reliance on physical post-polish measurements. The CMP machines under consideration are commercial polishing systems designed for 300 mm wafers. The process is a single-step polish operation with a commercially available polishing pad for oxide removal. The slurry used in the process is a commercially available formulation optimized for oxide CMP Table 1.

Aggregation of multivariate time series

The time series data (e.g., pressure, temperature, flow rate, and torque measurements) encompasses the sensor readings data recorded for the FDC. We employed basic aggregation functionals to compress the time series in a computationally efficient protocol. Table 4.1 presents the aggregation functionals applied for each sensor. According to the polishing protocol, each polishing operation is segmented into distinct polishing phases. The "start" step involves loading the wafer into the carrier, displacing water from the pad, and coating it with slurry. During the "ramp" step, the carrier is positioned on the pad, and target pressures and velocities are established. Finally, the "effective polish" step represents the primary polishing phase, which is stopped by in-situ optical endpoint detection. These steps are often succeeded by additional cleaning and conditioning of the pad and the platen.

In addition to the aggregation approach presented in this study, we also evaluated more advanced time-series feature extraction and comparison methods. For instance, we applied the `Tsfresh` library as suggested by Djedidi et al. (2022) to automatically generate a substantially larger set of statistical features and characteristics of the frequency domain per time series. We also tested kernel-based measures such as the Maximum Mean Discrepancy used by Cai et al. (2021) to compare all signal profiles. However,

neither approach improved predictive accuracy and both introduced significantly higher computational costs. This outcome can be attributed to the relative simplicity of many CMP sensor signals, which frequently exhibit near-constant or linear trends within defined polishing phases. For such signal structures, basic statistical aggregations (mean, std, auc, etc.) already capture the essential dynamics. Given the paramount importance of robustness in applications within a complex and high-stakes manufacturing environment, we posit that using statistical aggregates to extract features and a XGBoost model is a viable and pragmatic choice, as it balances predictive efficacy and operational simplicity.

Dataset overview

The dataset comprises 13,675 unique wafer runs from a technology node with one oxide polish step. The data stems from two CMP tools, each with multiple polishing chambers, and captures $\sim$ 6 million timestamps per polishing operation. The data spans over a time period of 13 months of productive process data. It involves two primary key performance indicators for VM: the spatially resolved MRRs and the derived WIWNU.

These metrics are determined by layer thickness measurements (spectroscopic reflectometry) from production control. We selected 17 distinct measurement sites for our dataset and numbered them radially in ascending order on the wafer. Exemplary positions can be seen in Fig. 4a. The positioning of the selected measurement sites depends on the specific product under investigation. The dataset includes over 30 different products. Since we numbered the chosen measurement sites on the wafer in ascending order, we can assign each measurement point to a corresponding wafer region. In this study, the wafer surface is therefore divided into five distinct zones, designated as Area1 (A1) through Area5 (A5) or "Edge region", as depicted in Fig. 4b. The 17-site measurement pattern is tailored to the specific geometry of each product type. Consequently, the absolute physical positions of the sites are product-dependent; however, the numbering scheme (sites 1-17) consistently corresponds to fixed relative locations, normalized from the defined areas. This approach allows for a spatially resolved

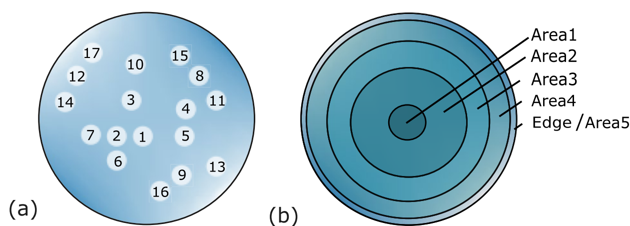

Fig. 4 **a** Illustrates an example set of selected measurement positions on the wafer surface. **b** Depicts the defined wafer regions in this study

comparison of property gradients across different product lines.

The change in layer thickness Δd at a given measurement point is defined as the difference between the pre- and post-polishing layer thickness measurements at that location. The material removal rate, which quantifies the polishing efficiency, is calculated as the change in layer thickness normalized by the effective polishing duration Δt,

$$\mathrm{MRR} = \frac{\Delta d}{\Delta t}. \tag{9}$$

To characterize localized thickness variations in a standardized manner, we employ the WIWNU metric. Following the definition by Luo and Dornfeld (2004), the WIWNU is computed as the difference between the maximum and minimum MRR values per wafer and per polishing operation, divided by twice the mean MRR

$$\mathrm{WIWNU} = \frac{MRR_{\max} - MRR_{\min}}{2 \cdot MRR_{\mathrm{avg}}} \times 100\%. \tag{10}$$

A perfectly smooth surface would result in a WIWNU of 0, since the maximal and the minimal difference of the spatially resolved MRRs would be 0. Hence, the greater the WIWNU, the worse the polish result.

The features of the dataset span over four categories. We include:

1. **Sensor readings**: Time series data (e.g., pressure, temperature, flow rate, and torque measurements)
2. **Process parameters**: Machine information and preprocess data
3. **Consumable data**: Usage counters for polish pads, retainer rings, and conditioners
4. **Derived features**: product-dependent offset factor (Equation 11)

(1) To compress the sensor time series data, the segments of the time series up to the "ramp" phase and including the primary polishing step are aggregated using the aggregation functionals described in Sect. 4.1 for each sensor. With this approach, we receive 348 distinct features from the time series sensor readings. To shrink down the number of aggregated features, we carefully reviewed the aggregations and used expert knowledge from Infineon engineers to remove non-relevant (e.g., "ramp" phase of the conditioner) or redundant features (e.g., different aggregations for near-constant features), resulting in 151 sensor-derived features.

(2) and (3) Process parameters and consumable counters stem from the same protocol and are subsequently included in the dataset. We conducted a thorough review

of the feature set and consulted with Infineon engineering experts to eliminate potentially confounding variables (e.g., operation index, fabrication facility identifiers) and redundant features (e.g., dresser lifetime correlated with cumulative wafer processing count), adding a total of 11 additional features to the dataset.

(4) Furthermore, a product-dependent MRR offset factor (*PF*) is introduced to capture the unique characteristics of each product type and their distinct surface structure. We discuss this factor in detail in the next section.

The resulting expert-selected dataset comprises 163 features spanning the four aforementioned categories. We incorporate the WIWNU as a single comprehensive target variable along with 17 selected measurement reference sites (MRRs) as target variables. These measurement sites are systematically assigned to defined wafer areas depending on radial distance from the wafer center.

Incorporating product identity

The product identity is not used as a categorical feature to maintain a favorable feature-to-sample ratio and to avoid modeling products as entirely independent classes. Instead, systematic product-to-product variation is captured through a multiplicative Product Factor (PF).

This approach is motivated by the nature of pattern-dependent effects in CMP. A product's specific layout characteristics, such as pattern density and surface topography, systematically influence polishing efficiency. These layout effects are not expected to add a constant offset to the removal rate but instead scale the overall polishing efficiency, acting as a product-specific proportional bias. The

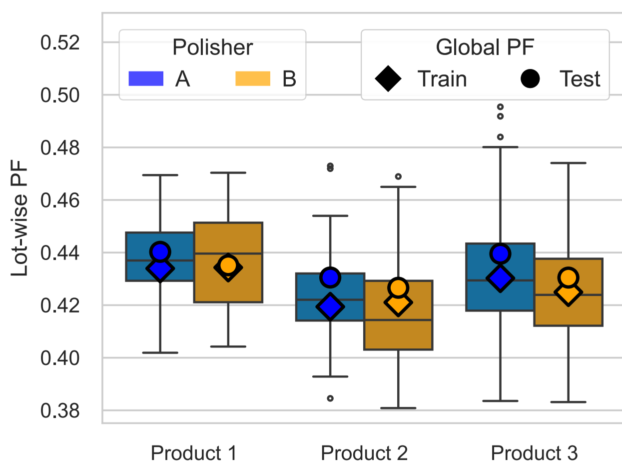

Fig. 5 Lot-wise product-dependent offset factors for the three most common products across the two CMP tools. Boxplots show the distribution with standard $1.5\times$IQR whiskers, with the central line indicating the median of the lot-wise factors. Diamonds and circles represent the global product-dependent offset factor computed on the complete train and test datasets, respectively

PF quantifies this bias by comparing a product's performance to a non-patterned baseline,

$$PF = \frac{MRR\,(\text{non-pattern wafer})}{MRR\,(\text{product})}. \tag{11}$$

Non-patterned (blank) wafers are test wafers with no patterned features or surface topography, used for daily tool monitoring.

In practice, this factor can be established in multiple ways. A global product-dependent offset factor is computed once over the full training dataset as the mean MRR of the non-patterned wafers divided by the mean MRR of the product wafers, then applied throughout. Alternatively, a lot-wise product-dependent offset factor is calculated dynamically by dividing the MRR of the closest-in-time non-patterned wafer by the median MRR of the product wafers in each lot.

To determine whether a simple global product-dependent offset factor is sufficient, a comparison between the two approaches was carried out. As an example, Fig. 5 shows the product-dependent offset factors for the three most common products on both machines, which together represent 44.31% of all wafer runs, ensuring statistical relevance for the analysis.

The factor is clearly product-dependent, as these products exhibit systematically different values. It is also tool-dependent, with consistent but small shifts observed between the two tools, confirming that calibration must be performed separately for each tool. With regard to temporal stability, the global factors computed on the train (diamonds) and test (circles) sets nearly overlap for every product-tool combination, demonstrating that the factor is less time-dependent. In addition, the global factors of training data consistently lie close to the lot-wise medians and within the interquartile ranges, demonstrating that it accurately captures the central tendency. Quantitative comparison supports this observation: differences between global and lot-wise means (not medians shown in the boxplots) are below 0.003 (<1%), and coefficients of variation are in the range of 3-5%, confirming small lot-to-lot variability and that the global factor provides a robust and practical approximation. Finally, the lot-wise factors were manually investigated for all products across the full dataset, showing no drift and no systematic shifts over time, such that continuous recalibration is not required. Nevertheless, periodic monitoring is advisable to account for potential changes. Based on these observations, the global factor from the training dataset was selected for subsequent modeling, providing a practical solution that minimizes calibration effort.

However, the approach does not directly generalize to previously unseen products. Any product absent from the dataset including the global PF requires an initial calibration

stage to obtain a reliable estimate of its product-specific mean MRR.

Model implementation

Our predictive modeling framework employs a structured pipeline, as depicted in Fig. 6, to analyze the prediction capabilities for the spatially resolved MRR measurements and the WIWNU models. The pipeline integrates:

- Data preprocessing (aggregation)
- Feature selection based on mutual information (Top 11, 20, 30 and 40 features retained)
- Model training with NCV
- SHAP investigation

The models for the spatially resolved MRR predictions are trained using the `MultiOutputRegressor` from sklearn. This implementation allows to create independent models for each output $\hat{y}_j$,

$$\hat{y}_j = \mathrm{Model}_j(\mathrm{x}), \quad j = 1, \dots, k. \tag{12}$$

Due to the spatial proximity of the 17 selected measurement sites, there is a potential risk of autocorrelation between adjacent points. To mitigate this issue, the predictive MRR models are trained independently.

To mitigate overfitting, the hyperparameter search space was constrained to the ranges depicted in Table 2 and we evaluated validation performance across the resulting hyperparameter sweeps.

The R^2 spread is calculated from the 10 best models of the outer loop of the NCV. The best hyperparameters from the inner loops are tracked and employed again to train final models with the complete training data. To avoid data leakage, they are tested with never-before-seen test data from a non-random time-based start split.

The final models are analyzed with the `TreeExplainer` from the SHAP library to provide SHAP values for each parameter entry in the data, see Sect. 5.3.

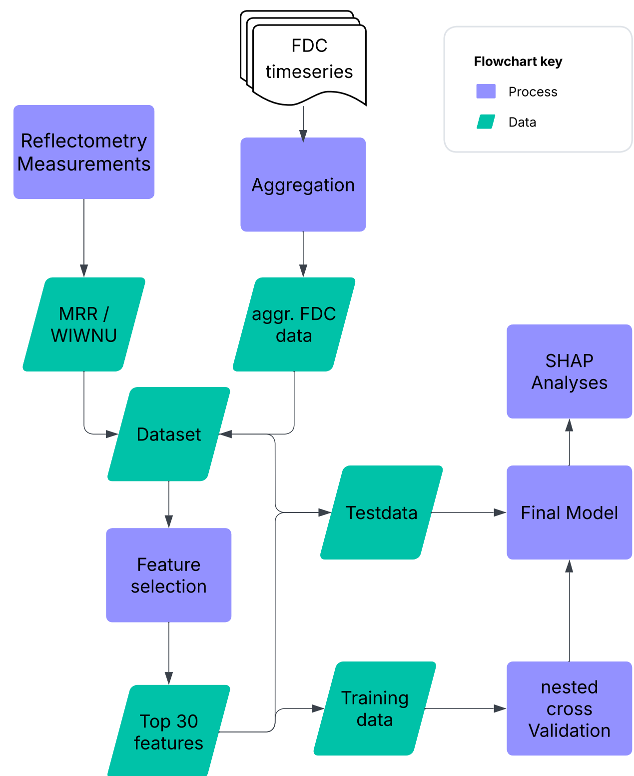

Fig. 6 Schematic representation of the machine learning workflow for spatially resolved MRR and WIWNU prediction

Table 2 Hyperparameter search ranges and utilized hyperparameters for XGBoost and RFR

Model	Hyperparameter	Lower bound	Upper bound	Best found
XGBoost	n_estimators	50	1000	800
	max_depth	2	12	5
	learning_rate	0.01	0.1	0.1
	subsample	0.7	0.9	0.7
	colsample_bytree	0.7	0.9	0.7
	gamma	0	0.2	0
RFR	n_estimators	50	100	50
	max_depth	10	20	20
	min_samples_split	2	10	5
	min_samples_leaf	1	4	2
	max_features	sqrt	None	sqrt

Results and discussion

Feature analysis

To evaluate the comparative performance of the time series aggregation approach for uniformity predictions, we conducted a sensitivity analysis across all aggregated representations of the platen rotation. Fig. 7 shows the strength of the linear correlations using a color gradient, where red denotes positive correlations and blue indicates negative correlations. The features extracted from different aggregation methods and sensor sources are arranged adjacent to facilitate direct comparison. The observed clustering patterns suggest that specific aggregation techniques exhibit a stronger linear influence on the target variable (WIWNU in the figure) than others, highlighting their relative importance in the predictive modeling framework.

Specifically, the effective polish step aggregations correlate strongly with post-polish non-uniformity. These aggregated features directly characterize the main polishing performance. Among them, the standard deviation (*std*) and

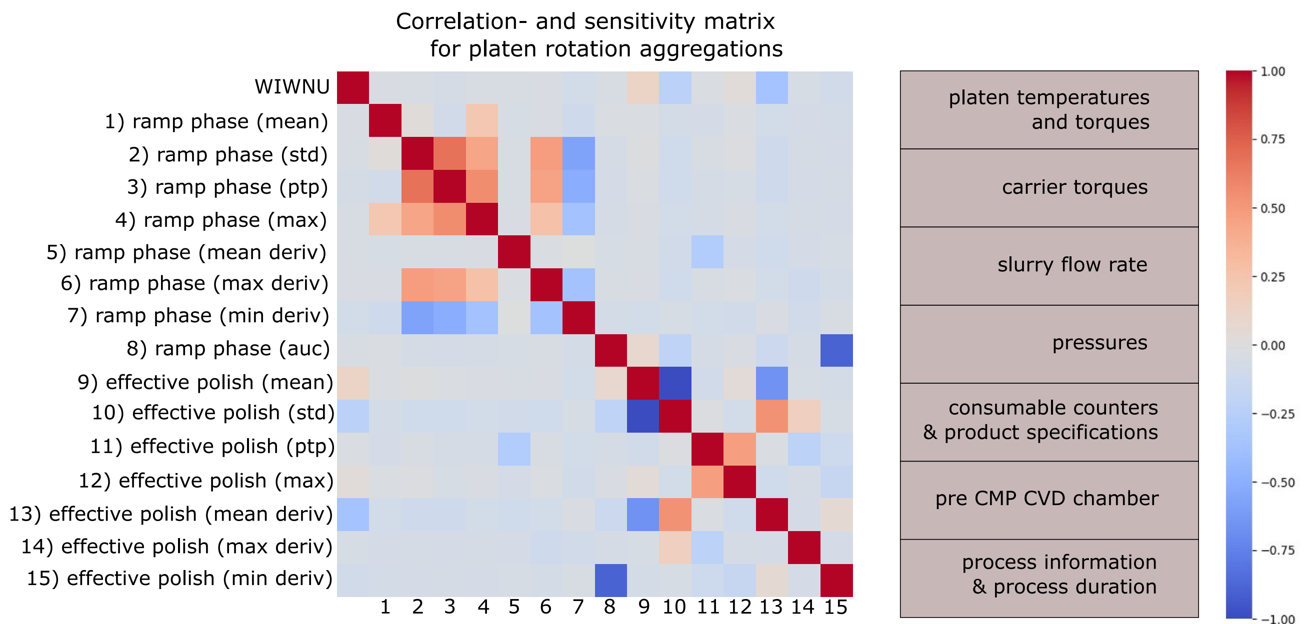

Fig. 7 Depicts a correlation- and sensitivity matrix with different derived aggregated features for the platen rotation. On the right, further data clusters of the expert-selected dataset are presented, which are aggregates as well

mean derivative (*mean deriv*) show a negative correlation with the WIWNU. These aggregation metrics quantify the ranges of signal fluctuation, providing critical insights into process stability. Given that the tool maintains key parameters like the platen rotation within narrow operational bounds, the tool's natural fine adjustments might benefit the uniformity result. The remaining aggregation methods positively correlate with the WIWNU, indicating a degradation in post-polish uniformity as these metrics increase. However, the sensitivity of this relationship varies depending on the specific aggregation method employed. Comparable trends are observed across other aggregated feature groups, as seen in Fig. 7.

The area under the curve (*auc*) aggregation method, which reflects cumulative trends over time, demonstrates a stronger correlation with the MRR for variables such as rotation, pressure and slurry flow rate. In contrast, the mean and peak-to-peak (*ptp*) aggregation methods significantly influence torque data, respectively. This phenomenon may be attributed to the functional interpretations of these aggregation methods: *ptp* captures the variations in a signal, while the *mean* reflects the average behavior.

Further cluster analysis indicates that the pressure data correlation is ambivalent. In general, it exhibits a slight positive correlation with removal rates, suggesting that higher pressures may improve removal efficiency. This corresponds to Preston's law, which states that the material removed is proportional to the mechanical work done on the wafer surface; thus, the MRR is proportional to the applied pressure.

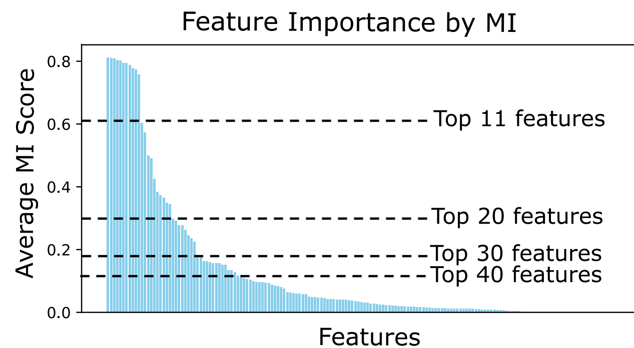

Fig. 8 Illustrates the MI scores for all features of the expert-selected dataset, sorted by MI score amount and the cutoff lines for the top 11, 20, 30, and 40 features. The top 5 features are the product-dependent offset factor, the carrier rotation (auc), the effective polish time, the platen rotation (auc), and the Area 3 pressure (auc)

Feature selection on the expert-selected dataset with 163 features was performed using mutual information, retaining the top 11, 20, 30, and 40 features that accounted for the most information gain, as subsequent features showed markedly diminished returns, see Fig. 8. These top features predominantly consist of data related to the platen, including torque measurements, rotational metrics, and temperature parameters. In addition, the selected features include consumable-related data, which capture critical aspects of consumable performance and wear behavior. Most of the remaining features originate from the aggregated time series data of the different pressure zones.

The feature analysis offers a foundation for predictive modeling, emphasizing the physical and operational aspects

of the CMP process that influence material removal and uniformity outcomes. In the following section, we assess the resulting VM models.

Model performances

LR is the baseline model, while RFR and XGBoost are employed as advanced ML algorithms to capture potential non-linear relationships in the data. The computational training costs of these models vary significantly, as summarized in Table 3. Their respective predictive performances are analyzed in detail in the following sections. The end-to-end single wafer run latency (feature aggregation + prediction) is 51 ms. These latencies are comfortably within the near-real-time requirements of the VM workflow.

Spatially resolved MRR predictions

The performance metrics of the three regression models, LR, RFR, and XGBoost, are evaluated using the average of the outer 10-fold

cross-validation. This section discusses the predictive capabilities of these models for the spatially resolved MRR predictions and the WIWNU predictions.

XGBoost emerges as the model that performs best in the in the evaluation of the 17 MRR measurement sites chosen, achieving the highest accuracy and the lowest performance variability, as shown in Table 4.

The XGBoost models demonstrate robust predictive capabilities even in the absence of the product-dependent offset factor. However, the inclusion of this factor significantly enhances the models' utility and accuracy in productive applications, underscoring its importance as a critical parameter for optimizing model performance.

Figure 9 presents the performance of the XGBoost model trained on the expert-selected dataset, showing the first 100 wafer runs from the test set. The figure shows the discrepancy between predicted and measured mean MRR (nm/s), with predictions averaged across the selected wafer areas (Areas 1-5) and contrasted against the corresponding area-averaged control measurements. The results demonstrate highly accurate predictions, particularly for the inner wafer areas. These regions exhibit minimal deviation between the predicted and actual MRRs, highlighting the model's robust performance.

Figure 10 compares ground truth and predicted MRRs across all 17 measurement sites, providing an overview of the model's overall predictive performance.

After the feature selection, the models achieve an average R^2 of $\mathbf{0.86 \pm 0.03}$, only relying on the 11 most important features, determined by their MI scores, including the PF. Incorporating 29 additional top features yields a marginally improved R^2 of $\mathbf{0.88 \pm 0.03}$, thereby indicating that the predictive efficacy of the model is adequately captured by the initial 11 top features.

Table 3 Compute times of the expert-selected models (4-16 core CPU, 2-4 GHz)

Model	Training time	Prediction time
LR	6 s per fold	1 ms
RFR	426 s per fold	10 ms
XGBoost	359 s per fold	13 ms

Table 4 Averaged performance comparison of the 17 MRR prediction models for different training sets

Model	Implementation	R^2	MAPE	95th pct. absolute error	99th pct. absolute error
XGBoost	expert-selected	0.89 ± 0.01	1.49%	0.36	0.52
	no PF included	0.83 ± 0.02	1.74%	0.37	0.58
	top 11 features	0.86 ± 0.03	1.78%	0.43	0.59
	top 20 features	0.87 ± 0.03	1.64%	0.39	0.56
	top 30 features	0.87 ± 0.03	1.62%	0.40	0.54
	top 40 features	0.88 ± 0.03	1.52%	0.39	0.54
	final model SHAP	0.86	1.63%	0.37	0.57
Other	LR (expert-selected)	0.83 ± 0.02	1.82%	0.49	0.66
	RFR (expert-selected)	0.83 ± 0.02	1.75%	0.46	0.66

Figure 11a and b demonstrate that near-optimal performance is achieved with comparatively few boosting rounds (n_estimators) and shallow trees (max_depth) for the expert-selected XGBoost models. Increasing model capacity beyond these values does not improve validation performance and, in this data-limited regime, is likely to exacerbate overfitting.

The sensitivity analysis shown in Fig. 11c was obtained by perturbing the PF variables by small increments around their nominal values in the expert-selected pre-trained models, with the objective of assessing feature stability. The results elucidate the potential consequences of minor calibration offsets encountered during the onboarding of previously unseen products. The best performance is achieved using an accurate PF, with a minimum MAPE of 1.5% and a R^2 close to **0.90**. Small negative deviations have a moderate effect, while positive misalignment above 10% leads to a sharp deterioration, with R^2 dropping below **0.5**. This highlights the importance of accurate offset calibration for the introduction of new products, as even slight overestimation can significantly impair predictive reliability.

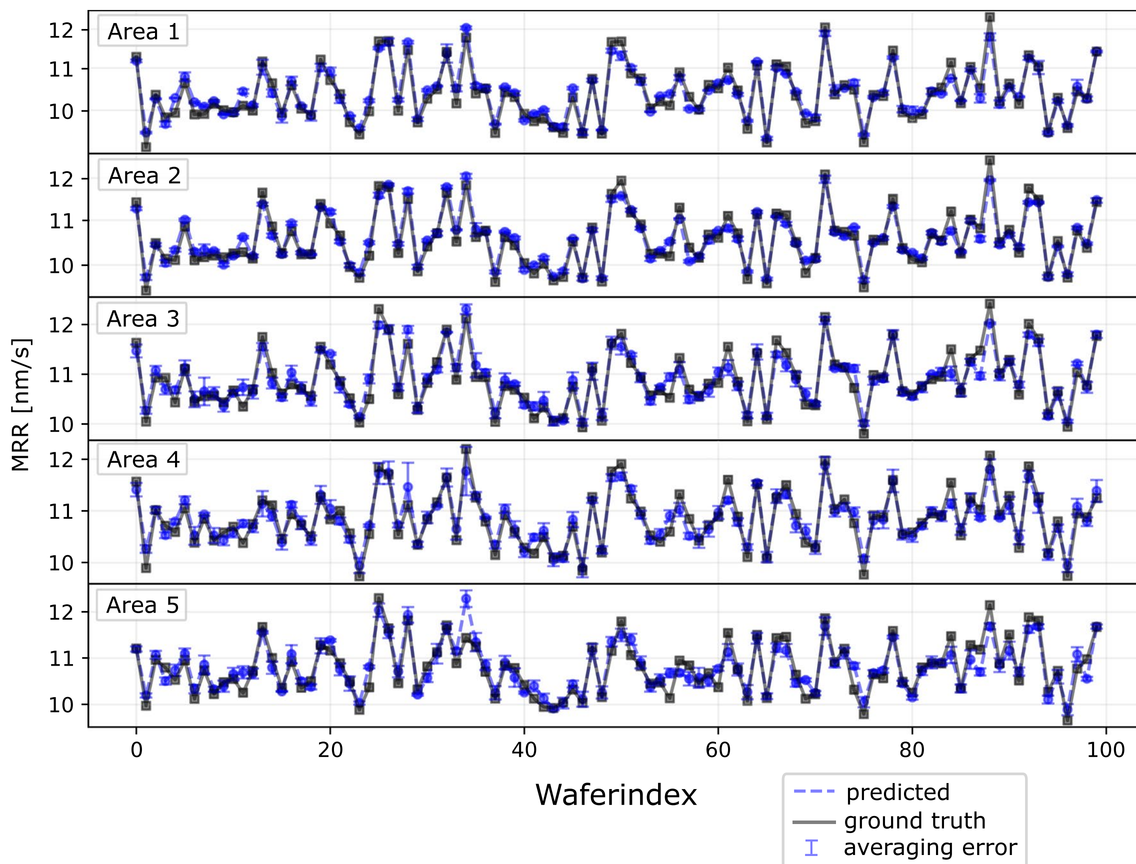

Fig. 9 Comparison of model performances for predicting spatially resolved MRRs using the expert-selected dataset with XGBoost. Predictions and ground truth values are averaged across the chosen measurement sites according to their respective wafer surface areas. Error bars represent the site-averaged error

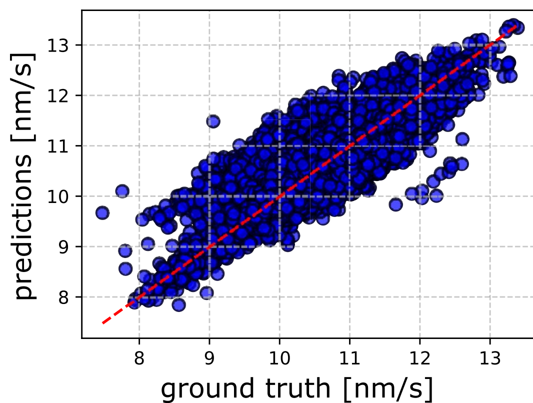

Fig. 10 Ground truth vs. predicted MRR scatter plot for all 17 sites with a 1:1 reference line for the expert-selected dataset with XGBoost

A smaller-than-accurate PF leads the models to assume a higher product MRR, which results in a systematic underestimation of removal. These errors remain moderate, as the predictions stay within a physically plausible range. In contrast, an overestimated factor makes the models assume a lower product MRR, thereby systematically over-predicting removal. This effect could compound, particularly in regions of high removal, potentially leading to a steeper performance decline with positive PF deviations.

The specification limits for the MRR in the industry reported by Infineon engineers typically range between $\pm 0.3\ \mathrm{nm/s}$ and $\pm 0.8\ \mathrm{nm/s}$, depending on the process layer and technology node. We therefore evaluate all models as capable of substituting the post-CMP metrology measurements in production. Across the tested conditions, the wafer level model predictions remain within the specification limits defined by industry standards. As summarized in Table 4, the wafer-level 99th-percentile absolute prediction error, used as a proxy for worst-case production risk, remains small. We discuss further possible improvements of the models in Sect. 6.

WIWNU prediction

The LR WIWNU prediction model cannot capture the underlying data structure, demonstrating significant constraints in predictive performance. In contrast, the RFR and XGBoost models substantially improve predictive accuracy.

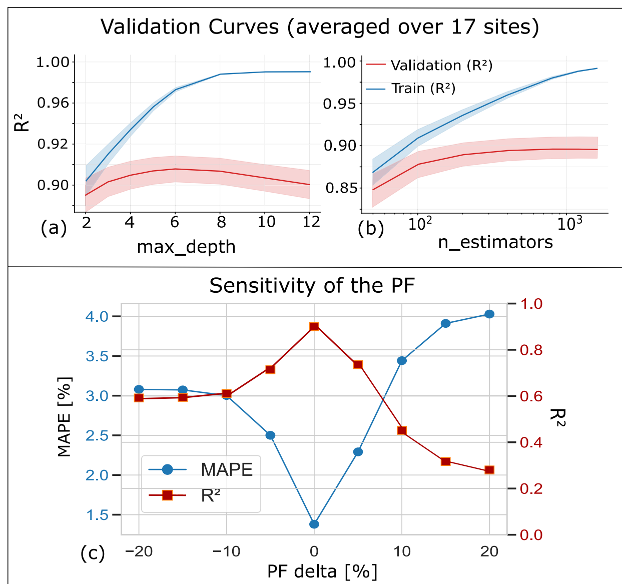

Fig. 11 The figure illustrates a multifaceted analysis of performance metrics for the XGBoost model and the expert-selected dataset. Subplots (a) and (b) depict the validation curves resulting from the extension of hyperparameters. Subplot (c) presents a sensitivity analysis of the R^2 and MAPE metrics with respect to perturbations in the product-dependent offset factor, revealing that even minor adjustments to this parameter in the test set can significantly influence the predictive efficacy of the model

Table 5 Performance comparison of WIWNU prediction for various models and training sets

Model	R^2	MAPE
LR (expert-selected)	0.32 ± 0.03	17.74%
RFR (expert-selected)	0.65 ± 0.03	12.85%
XGBoost (expert-selected)	0.64 ± 0.09	12.78%
XGBoost (no PF included)	0.56 ± 0.15	16.71%
XGBoost (top 30 features)	0.59 ± 0.19	15.53%
XGBoost (final model SHAP)	0.72	11.24%

However, their mean predictive accuracy remains inferior to that of spatially resolved MRR prediction models.

The exclusion of the PF leads to a measurable degradation in model performance, underscoring its significance for accurate post-polish non-uniformity prediction. Subsequent feature reduction to the 30 most important features yields comparable predictive accuracy when the PF is included.

All WIWNU prediction models have a relatively high R^2 spread over the 10 folds. Given the performance results from Table 5, we evaluate the models as not yet applicable for use in productive wafer manufacturing . However, key improvements can still be implemented and are discussed in Sect. 6.

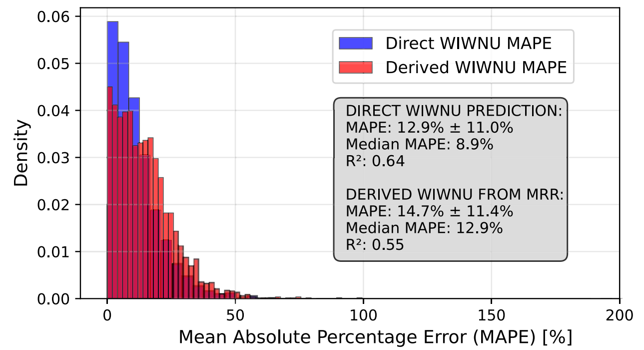

Fig. 12 Error distributions of direct WIWNU predictions (blue) and WIWNU derived from 17 predicted MRRs (orange). While the direct model exhibits narrower and more concentrated errors, the derived WIWNU suffers from amplified deviations due to its dependence on predicted extreme MRR values, resulting in a broader distribution

Comparison of direct and derived WIWNU predictions

Another option to predict WIWNU is to calculate it from the 17 individual MRRs predicted by the best XGBoost model trained on the expert selected dataset as described in Sect. 5.2.1. This results in a R^2 of **0.55**, when compared to the measured WIWNU, which corresponds to the WIWNU models performance and demonstrates no substantial improvement. The reliance of the WIWNU formula on maximum and minimum MRR values may render it sensitive to large errors, whereas the direct WIWNU model learns to be robust against such failures. This suggests that the directly trained WIWNU model captures aspects of the data not represented by the simple max-min calculation. The cumulative effect of individual, smaller-scale errors emanating from the 17 distinct predictive models may result in a significant, aggregated error in the WIWNU prediction, further compromising the accuracy of the overall WIWNU prediction from the 17 predicted MRRs.

To quantitatively validate the robustness of direct WIWNU modeling, a comparison against WIWNU values derived from the predicted 17-sites MRR distributions was done. The histogram of the MAPE distributions in Fig. 12 confirms the consistent advantage of the direct model across the test dataset. Moreover, the derived predictions display heavier-tailed error distributions, indicating that small deviations in predicted minimum and maximum MRR values propagate into disproportionately large WIWNU errors.

In the next section we apply SHAP analyses for post hoc explainability.

Model explanation

Explanation of the WIWNU prediction

First, the overall trend of feature importance in non-uniformity predictions is discussed, and subsequently, the feature importance of the ML models for the prediction of the MRR for specific measurement sites is examined. The SHAP summary plot in Fig. 13 visualizes the 10 most impactful features, according to the mean absolute SHAP values across all expert-selected XGBoost model samples. In the SHAP plot, each data point corresponds to the SHAP value (x-axis) of a feature for an individual observation, indicating its directional impact on the model's output (positive impact is on the right, negative impact is on the left). The color gradient encodes the relative magnitude of the original feature value. Red dots represent high feature values, while blue dots indicate low ones.

The **product-dependent offset factor** is the most influential feature. It quantifies inherent material removal variations across pattern densities and structures (e.g., dense vs. sparse features). If certain products exhibit consistent removal rate deviations across the wafer, their cumulative impact can influence WIWNU. For example, if dense patterns systematically polish slower, the average effect may dominate the WIWNU. The maximum SHAP contributions reach approximately $\pm 1.5\,\%$ change in WIWNU. Relative to the mean WIWNU of our dataset of $4.3\,\%$, these effects account for about 29 % of the mean variation, indicating controllable sensitivities.

The **edge pressure** (*std*) is the second most influential feature of the model, suggesting a strong correlation between Edge control and the resulting WIWNU. The *std* metric, which represents the range of fluctuations in the pressure over time, measures the process's dynamic stability. Since pressure fluctuations are extremely small, minimal

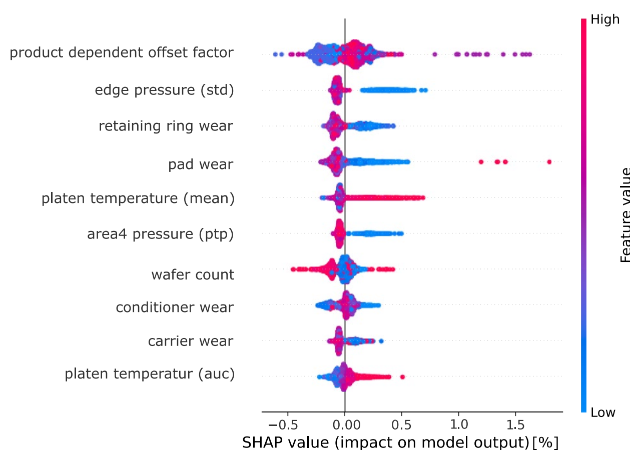

Fig. 13 Illustrates the SHAP summary plot for the 10 most important features of the final XGBoost model for predicting the WIWNU

adjustments in rotation speed can benefit the post-polishing WIWNU result, as suggested by the SHAP analyses.

Consumable-related features (**retaining ring wear**, **pad wear**, **conditioner wear**) exhibit a strong influence on the WIWNU prediction, highlighting possible first wafer effects. Higher consumable wear is correlated with decreased post-polish WIWNU. Wafers with freshly applied consumables might experience "break-in" effects, while later wafers benefit from stabilized consumables. The **wafer count** is a proxy for the tool wear, exhibiting a positive correlation with increasing non-uniformity. This observation suggests potential tool-related aging effects.

The SHAP analysis reveals that the WIWNU is predominantly governed by inherent pattern-dependent removal variations and modulated by consumable-related features. However, the relatively wide spread of the performance metrics (R^2) and the relatively small model performance overall should be considered when interpreting the model and drawing conclusions. In the following section, the spatially resolved MRRs model is investigated. It can be more reliably interpreted because of the relatively strong performance.

The **platen temperature** shows a positive relation, suggesting that thermal expansion or altered slurry dynamics exacerbate non-uniformity. The **Area4 pressure** (*ptp*) metric mainly expresses the pressure fluctuation in the defined area.

Explanation of the spatially resolved MRR predictions

Figure 14 shows four exemplary SHAP summary plots, illustrating the 10 most influential features of the final XGBoost models in predicting the 17 spatially resolved MRRs. The plots correspond to the measurement sites 4, 9, 13, and 17, which we described with varying radial distances from the center of the wafer.

A visual examination of the SHAP summary plots reveals a degree of variability in the most important features across the different models. However, certain features consistently emerge as key contributors, including the **effective polish time**, **carrier rotation** (*mean derive*), **product-dependent offset factors**, and the **slurry flow rate**. The recurrence of these features across multiple models suggests their significance in predicting the MRRs, warranting further investigation into their underlying relationships.

The **effective polish time** emerges as a critical predictor, especially for the inner wafer areas. With increasing effective process duration Δt, the MRR is decreased. Since the process is terminated via in-situ optical endpoint measurements, it is not regulated by fixed polish times. Equation 9 indicates that with increasing Δt the removal rate decreases, while layer thickness is kept constant. This effect is reflected

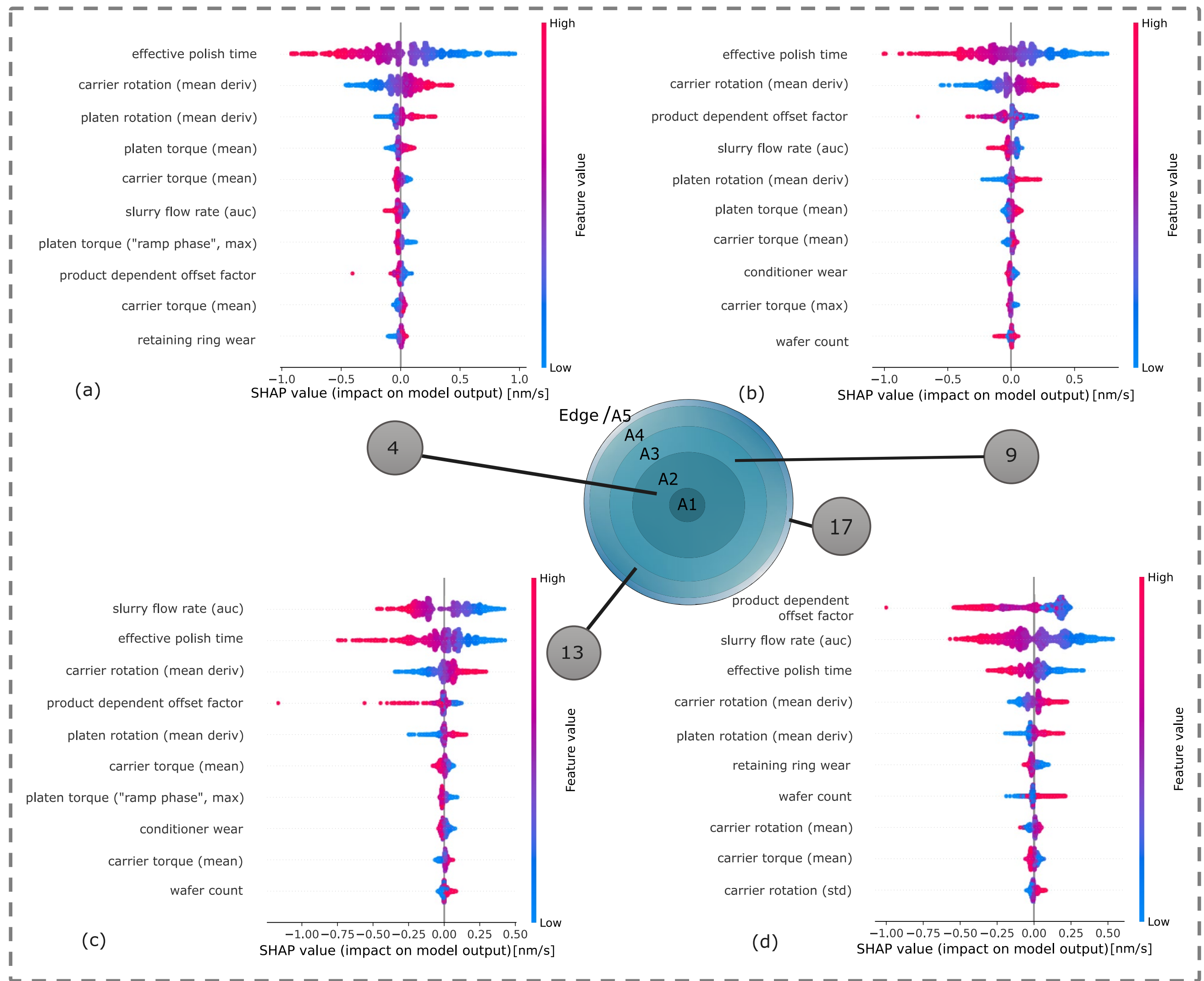

Fig. 14 Shows the SHAP summary plot for the 10 most important features for the MRR prediction XGBoost models. The numerical values denoted by the grey circle labels correspond to relative radial positions across the wafer surface with **a** Site 4 (near-center), **b** Site 9 (mid-wafer region), **c** Site 13 (outer-wafer region), and **d** Site 17 (wafer edge)

in the SHAP plot, where increasing Δt decreases the MRR prediction.

The **effective polish time** exhibits SHAP values up to approximately $+0.9$ nm/s, corresponding to 1.8 times the allowed process deviation (typical range between ± 0.3 and ± 0.8 nm/s,). This means that deviations in polish duration alone can push the predicted MRR beyond the acceptable process window, emphasizing the critical role of endpoint control. Similarly, the mean derivatives of **carrier rotation** and **platen rotation** show maximum contributions of about $+0.45$ nm/s and $+0.35$ nm/s, respectively, each approaching or exceeding 70-90% of the tolerance band. When these dominant parameters act simultaneously in the same direction, their combined SHAP contributions can sum to approximately 1.7 nm/s, exceeding the specified process window.

High **product-dependent offset factors** correlate with negative values (blue dots, right side). This suggests that certain products inherently polish with a smaller MRR due to their removal rate deviations from the blank disc reference. Low offset values (blue dots, left side) are associated with negative SHAP values, implying lower MRR for products closer to the blank disc's removal behavior. This effect seems especially important in the Area4 and Area5 regions of the wafer, showing maximum contributions of about $+0.40$ nm/s and -1 nm/s, approaching over 100% of the tolerance band in rare cases.

Rotation and torque metrics of the platen and the carrier exhibit significant importance across all SHAP plots. In particular, the *mean derivative* and *mean* aggregation functionals are consistently identified as influential features. The *mean derivative* of the rotation and torque metrics captures

short-term fluctuations in rotational speed under near-constant RPM conditions, whereas the *mean* represents the steady-state behavior of the system. Positive changes in the *mean derivative* metrics of the rotation and torques seem to increase the MRR prediction, further underscoring that minimal adjustments of the tool help to improve the polishing result.

Slurry **flow rate** *auc* emerges as a critical predictor for the Area4 and Area5 MRR models. Low/erratic **flow rate** *auc* (blue dots, right SHAP values) correlates with higher material removal rates (positive SHAP), and high **flow rate** *auc* (red dots, left SHAP values) decreases MRRs (negative SHAP). Although the common operational intuition is that more slurry provides higher MRR, our SHAP-based analysis indicates a more nuanced relationship. The observed correlation is plausibly driven predominantly through the AUC of the flow-rate signal. Because the flow rate is approximately constant, the AUC scales primarily with process duration rather than true variability in flow. Achieving on-target polishing requires extended process time and thereby reduced MRR. Slurry **flow rate** *auc* contribute less than $0.1\ \mathrm{nm/s}$ on average, which is below 20% of the spec tolerance.

The SHAP analysis reveals that temporal parameters and rotational process control dominate spatially resolved MRR predictions. Although feature importance varies across individual measurement sites, these key factors consistently influence all models at the selected sites.

These findings provide mechanistic insight into the role of machine components in the process, with further implications discussed in the conclusion and outlook section.

Comparison to existing CMP VM studies

When comparing our results to the literature, it becomes clear that prior CMP VM studies differ substantially in dataset, methodology, and target definition, as shown in Sect. 2. Comparisons are limited to studies related to MRR, as to our knowledge, no prior work addresses the prediction of WIWNU.

Tsuda et al. (2015) identified average pressure values, slurry flow, and motor current as key predictors of removal rate using partial least squares followed by multiple linear regression. This agrees with our findings, highlighting pressure distributions, slurry flow, and mechanical load (torque and rotation) as key drivers. The main differences are that their analysis is restricted to a single globally averaged MRR and linear effects, whereas our models extend to spatially resolved predictions and capture nonlinear dependencies. More recent work on the public PHM dataset, typically using random splits, highlights a different issue. Zuo et al. (2024) applied SHAP with XGBoost and found dresser table usage as the most important feature, followed by pressure, slurry flow, and rotation. Li et al. (2019) used RFR to find slurry flow variation and pressure skewness among the top drivers, but also usage-related variables such as backing film and table usage. Zhang et al. (2021) and K. Chen et al. (2024) both relied on Random Forest feature importance as well and consistently ranked usage counters (dresser, table, carrier, backing film, membrane) as the most significant features. These usage variables increase monotonically with wafer chronology and are therefore prone to temporal leakage, raising concerns about the physical interpretability of such models. In contrast, our analysis on large-scale industrial data with a time-based split reveals physically meaningful drivers such as slurry flow, torque, and pressure distributions, and extends the scope from global wafer averages to spatially resolved MRR and, for the first time, WIWNU prediction.

Conclusion and outlook

The proposed models were designed to use derived FDC features and process information to predict post-polish non-uniformity and spatially resolved MRR results. We demonstrated the potential of using ML techniques for this task, enabling process control improvements and reducing reliance on physical wafer measurements. The main findings of this study are listed below.

(1) Spatially resolved MRRs can be predicted even with statistical aggregation functionals to compress sensor time series with high precision using XGBoost.
(2) We found that the *std* and *mean* aggregation of the sensor time series provided the most valuable contribution to the model's prediction capabilities. It captures deviations and the system's steady state behavior, respectively.
(3) A post hoc explanation with SHAP provides information on the CMP process and possible process adjustments. We found that especially polishing time and rotation metrics govern the spatially resolved MRR predictions. Models for different measurement sites on the wafer surface exhibit different feature relevance.

The main current limitations of this study are listed below.

(1) The WIWNU prediction model shows promising results but requires further refinement to achieve application readiness. A hybrid physics-ML route can materially lift accuracy and robustness. One effective pattern could be to encode chamber and wafer physics as a differentiable, low-order "grey-box" layer and let ML learn the residuals. One could parameterize this physics layer by

recipe settings and tool identifiers to capture systematic structure.

(2) The importance of a baseline factor that differentiates products and their surface structures is underscored, especially showing the value in a high-mix manufacturing environment. However, when included, each new product must undergo a calibration phase in which the mean MRR for the newly added product must be determined. To improve practical applicability, the key is to replace the PF with physically meaningful design features, adopt transfer learning and domain adaptation so the model learns product-invariant structure, and add lightweight, adaptive calibration mechanisms that require only a few wafers to “warm start” a new product.

(3) Aggregating the sensor reading time series data by applying statistical aggregators might lead to a loss of information throughout the complex process.

Incorporating endpoint signal-derived features, which, depending on the specific process, may be based on optical or motor current measurements could improve the models. Features derived from *in-situ* reflectometry measurements used for optical endpoint detection could provide spatially resolved, real-time removal rate information. Other CMP processes, such as poly-Si polishing, may benefit from including motor current time series in the dataset. Since the deceleration profile of the motor current upon barrier contact can serve as an indicator of *in-situ* wafer uniformity, its inclusion may enhance model predictive accuracy. Finally, the time series data aggregation could be optimized by employing sensor-specific aggregation strategies. This approach would maximize information extraction from each time series while minimizing the number of feature parameters, thus improving data fidelity without sacrificing computational efficiency.

Beyond prediction, the trained models also offer potential for prescriptive analytics. Since they approximate non-linear mapping between process settings and spatially resolved MRRs with high fidelity, they can be repurposed as surrogate simulators in optimization frameworks. For example, together with Bayesian optimization or evolutionary algorithms, they could enable automated search for process recipes that minimize WIWNU or achieve specific target profiles. This inverse optimization of process parameters would extend the utility of virtual metrology from passive prediction to active recipe optimization, providing a natural next step toward autonomous CMP control.

The spatially resolved MRR prediction model can be directly integrated into industrial CMP processes, supporting the semiconductor industry’s transition toward data-driven and cost-efficient manufacturing. The models, as shown, can reduce the amount of metrology measurements and improve process control in the future.

Acknowledgements The authors gratefully acknowledge Andre Pohl and Olaf Kühn (Infineon Technologies Dresden) for their essential contributions to the project’s preparation and their valuable guidance throughout its execution. The authors sincerely thank the European Regional Development Fund (EFRE) and the Free State of Saxony of the Federal Republic of Germany for their funding of this work.

Author Contributions *Conceptualization*: Morten Breidung; *Methodology*: Morten Breidung, Manuel Guenther, Tom Rothe; *Formal analysis and investigation*: Morten Breidung, Tom Rothe, Manuel Guenther, Andre Lauff, Peter Thieme; *Writing - original draft preparation*: Morten Breidung; *Writing - review and editing*: Morten Breidung, Tom Rothe, Andre Lauff, Jan Langer; *Funding acquisition*: Peter Thieme, Harald Kuhn, Tom Rothe; *Supervision*: Harald Kuhn, Jan Langer, Andre Lauff, Peter Thieme.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The raw data are confidential production data from an Infineon fab and are protected under a cooperation agreement, ensuring responsible handling and safeguarding of proprietary information. Therefore, sharing the data is not possible.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relations that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Breiman, L. (2001). *Random forests. Machine learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Cai, H., Feng, J., Yang, Q., Li, F., Li, X., & Lee, J. (2021). Reference-based virtual metrology method with uncertainty evaluation for material removal rate prediction based on gaussian process regression. *The International Journal of Advanced Manufacturing Technology, 116*(3), 1199–121. https://doi.org/10.1007/s00170-021-07427-2

Chen, K., Xiao, B., Liu, X., Wang, C., & Liang, S. (2024). Bidirectional long-short term memory predictor for material removal rate in computer-controlled optical surfacing. *Precision Engineering, 89*, 473–49. https://doi.org/10.1016/j.precisioneng.2024.07.006

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794https://doi.org/10.1145/2939672.2939785

Cho, Y., Kim, M., Hong, M., Han, J., Kim, H. J., Kim, H., & Lee, H. (2025). Prediction of normalized material removal rate profile based on deep neural network in five-zone carrier head cmp system. *International Journal of Precision Engineering and Manufacturing-Green Technology, 1–1*, https://doi.org/10.1007/s40684-025-00698-0

Deivendran, B., Masampally, S., Nadimpalli, V., & Runkana, V. (2025). Virtual metrology for chemical mechanical planarization of semiconductor wafers. *Journal of Intelligent Manufacturing, 36*(3), 1923–1942. https://doi.org/10.1007/s10845-024-02335-0

Djedidi, A., Clain, S., Borodin, V., & Roussy, G. (2022). Feature selection for virtual metrology modeling: an application to chemical mechanical polishing. *2022 33rd Annual SEMI advanced semiconductor manufacturing conference (ASMC)*, 1–6https://doi.org/10.1007/s00170-021-07427-2

Farahani, M. A., Kalach, F. E., Harper, A., McCormick, M., Harik, R., & Wuest, T. (2025). Time-series forecasting in smart manufacturing systems: An experimental evaluation of the state-of-the-art algorithms. *Robotics and Computer-Integrated Manufacturing, 95*, 10301. https://doi.org/10.1016/j.rcim.2025.103010

Jebri, M.A., Graton, G., El Adel, E., Ouladsine, M., & Pinaton, J. (2016). Virtual metrology on chemical mechanical planarization process based on just-in-time learning. *2016 5th International Conference on Systems and Control (ICSC)*, 169–174https://doi.org/10.1109/ICoSC.2016.7507082

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E, 69*, Article 066138. https://doi.org/10.1103/PhysRevE.69.066138

Lee, K. B., & Kim, C. O. (2018). 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

Li, Z., Wu, D., & Yu, T. (2019). Prediction of material removal rate for chemical mechanical planarization using decision tree-based ensemble learning. *Journal of Manufacturing Science and Engineering, 141*(3), 03100. https://doi.org/10.1115/1.4042051

Liu, C.-L., Tseng, C.-J., Hsaio, W.-H., Wu, S.-H., & Lu, S.-R. (2022). Predicting the wafer material removal rate for semiconductor chemical mechanical polishing using a fusion network. *Applied Sciences, 12*(11), 11478. https://doi.org/10.3390/app122211478

Lundberg, S.M., Erion, G.G., & Lee, S.-I. (2018). *Consistent individualized feature attribution for tree ensembles.* https://doi.org/10.48550/arXiv.1802.03888. ArXiv.

Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774).

Luo, J., & Dornfeld, D. (2004). Wafer-scale modeling of cmp. In: *Integrated modeling of chemical mechanical planarization for sub-micron ic fabrication: From particle scale to feature, die and wafer scales* (pp. 255–284). Springer.

Prognostics and Health Management Society (2016). *PHM Data Challenge 2016.* https://phmsociety.org/conference/annual-conference-of-the-phm-society/annual-conference-of-the-prognostics-and-health-management-society-2016/phm-data-challenge-4/. (Accessed: 2025-03-10)

Rahman, M. W., Vogl, G. W., Jia, X., & Qu, Y. (2024). Physics-informed multi-task learning for material removal rate prediction in semiconductor chemical mechanical planarization. *IEEE International Conference on Prognostics and Health Management (ICPHM), 2024*, 385–39. https://doi.org/10.1109/ICPHM61352.2024.10627679

Rothe, T., Lauff, A., Shaporin, A., Thieme, P., Sayyed, M.A., Gottfried, K. & Stoll, M. (2025). Real-time interfacial pressure prediction in cmp using machine learning surrogates of finite element simulations. *International Journal of Automation Technology*, *19*(5), 879–88https://doi.org/10.20965/ijat.2025.p0879

Shapley, L., & (1997). A value for n-person games. contributions to the theory of games ii. (1953). 307–317. In H. W. Kuhn (Ed.), *Classics in game theory* (pp. 69–79). Princeton: Princeton University Press.

Shauly, E., & Rosenthal, S. (2020). Coverage layout design rules and insertion utilities for cmp-related processes. *Journal of Low Power Electronics and Applications, 11*(1), 2. https://doi.org/10.3390/jlpea11010002

Shiul, S.-J., Yu, C.-C., Shen, S.-H., & Sul, A.-J. (2004). Multivariable control of multi-zone chemical mechanical polishing. *Proceedings of the 2004 Semiconductor Manufacturing Technology Workshop*, 107–110https://doi.org/10.1109/SMTW.2004.1393737

Tsuda, T., Inoue, S., Kayahara, A., Imai, S.-I., Tanaka, T., Sato, N., & Yasuda, S. (2015). Advanced semiconductor manufacturing using big data. *IEEE transactions on semiconductor manufacturing, 28*(3), 229–235. https://doi.org/10.1109/TSM.2015.2445320

Wang, T., Lu, X., Zhao, D., & He, Y. (2013). Contact stress non-uniformity of wafer surface for multi-zone chemical mechanical polishing process. *Science China Technological Sciences, 56*, 1974–197. https://doi.org/10.1007/s11431-013-5245-y

Wei, Y., & Wu, D. (2022). Material removal rate prediction in chemical mechanical planarization with conditional probabilistic autoencoder and stacking ensemble learning. *Journal of Intelligent Manufacturing, 35*(1), 115–127. https://doi.org/10.1007/s10845-022-02040-w

Winkler, G., Rothe, T., Sayyed, M. A., Jäckel, L., Langer, J., Kuhn, H., & Stoll, M. (2025). Machine learning in chemical-mechanical planarization: A comprehensive review of trends, applications, and challenges. *Advanced Engineering Informatics, 68*, Article 103663. https://doi.org/10.1016/j.aei.2025.103663

Xia, L., Zheng, P., Huang, X., & Liu, C. (2021). A novel hypergraph convolution network-based approach for predicting the material removal rate in chemical mechanical planarization. *Journal of Intelligent Manufacturing, 33*(8), 2295–2306. https://doi.org/10.1007/s10845-021-01784-1

Yi, J., Sheng, Y., & Xu, S. (2003). Neural network based uniformity profile control of linear chemical-mechanical planarization. *IEEE transactions on semiconductor manufacturing, 16*(4), 609–62. https://doi.org/10.1109/TSM.2003.818987

Zhang, J., Jiang, Y., Luo, H., & Yin, S. (2021). Prediction of material removal rate in chemical mechanical polishing via residual convolutional neural network. *Control Engineering Practice, 107*, 10467. https://doi.org/10.1016/j.conengprac.2020.104673

Zuo, J., Chen, Z., Cui, Y., Cheng, Y., Ma, Y., Chen, H. & Gao, D. (2024). Predicting material removal rate in chemical mechanical polishing (cmp) using explainable machine learning methods. *Proceedings of the 2024 Conference of Science and Technology for Integrated Circuits (CSTIC)*, 1–https://doi.org/10.1109/CSTIC61820.2024.10532078

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