Sensors and Materials, Vol. 38, No. 4 (2026) 2085–2096
MYU Tokyo

S & M 4424

Machine-learning-based Solution for Thin-film Quality in Chemical Vapor Deposition Coating Processes

Shu-Han Liao,[1*] Cheng-Yu Tsai,[2] Bo-Ruei Su,[1]
Wei-Jhong Chen,[1] Yu-Wei Li,[1] and Li-Sheng Chen[3**]

[1]Department of Electrical and Computer Engineering, Tamkang University
No. 151, Yingzhuan Rd., Tamsui Dist., New Taipei City 251301, Taiwan
[2]Department of Business Administration, National Taiwan University of Science and Technology
No. 43, Sec. 4, Keelung Rd., Da'an Dist., Taipei City 106335, Taiwan
[3]Department of Computer Science and Information Engineering, National Ilan University
No. 1, Sec. 1, Shennong Rd., Yilan City, Yilan County 260007, Taiwan

(Received June 5, 2025; accepted April 1, 2026)

Keywords: semiconductor coating process, machine learning, feature selection, parameter optimization, chemical vapor deposition

In this study, we developed a data-driven method to optimize thin-film processes in semiconductor manufacturing by integrating chemical vapor deposition with machine learning. By analyzing 91 process parameters and applying the analysis of variance and principal component analysis, we identified key variables influencing photoluminescence quality. A predictive model was built using the random forest (RF) algorithm and compared with the k-nearest neighbors method. The RF-based model demonstrated superior accuracy and robustness. The proposed method improves process stability, increases yield, and supports automated intelligent manufacturing across diverse material systems. Quantitatively, the RF-based configurations achieved R^2 as high as 0.936 (mean intensity) and 0.842 (wavelength standard deviation (STD)), with mean squared error (*MSE*) minimized to 2.71×10^6 for mean-intensity prediction and 2.064 for wavelength-STD prediction, corresponding to up to a 96.7% MSE reduction relative to the k-nearest neighbors (KNN) baseline.

1. Introduction

Thin-film deposition is a key stage in semiconductor production because coating quality directly affects device behavior and manufacturing yield. In practice, recipe tuning is still often adjusted by engineers experienced and repeated trial-and-error, which becomes inefficient when the process window is narrow or when new materials are introduced.(1)

To address this issue, in this work, we combined CVD process data with machine learning analysis and examined 91 process variables, used analysis of variance (ANOVA) and principal component analysis (PCA) to screen influential factors, and then trained predictive models with

*Corresponding author: e-mail: shliao@gms.tku.edu.tw
**Corresponding author: e-mail: lschen@niu.edu.tw
https://doi.org/10.18494/SAM5779

ISSN 0914-4935
https://myukk.org/

k-nearest neighbors (KNN) and random forest (RF) to estimate thin-film quality and assist recipe adjustment.

This scheme offers a data-based route for CVD process optimization in semiconductor manufacturing. By linking key parameters to quality indicators, it can help improve deposition consistency, support yield enhancement, and provide a basis for adaptive and automated control in different material systems.

2. Related Works

Machine learning (ML) techniques have been widely applied in various fields for feature extraction, dimensionality reduction, and process optimization. Fang *et al.* proposed a feature selection method for synthetic aperture radar (SAR) imagery using PCA, demonstrating that PCA effectively removes noise and enhances boundary preservation.[2] Although their work focuses on image data, the underlying concept of using PCA to extract dominant features is directly applicable to high-dimensional manufacturing data such as those in CVD processes.

In semiconductor manufacturing, data-driven "virtual metrology" (VM) has been widely studied to estimate film characteristics (e.g., thickness) without time-consuming offline measurements. In early VM works, statistical learning frameworks were proposed to predict CVD thickness and practical constraints such as latency and real-time deployment considerations were discussed.[3,4] For plasma-enhanced processes, plasma/sensor information has been used to build VM models for multilayer plasma-enhanced chemical vapor deposition (PECVD) nitride thickness prediction.[5] Recent surveys further summarize VM development trends and highlight challenges such as sensor drift, feature selection, and model interpretability for fab adoption.[6] Beyond VM, comparative studies in semiconductor process monitoring have benchmarked combinations of feature extraction/selection and classifiers to clarify which algorithms are robust under different fault scenarios.[7] In terms of experimental design and analysis, standard design of experiments (DOE)/ANOVA frameworks remain mainstream tools to quantify factor effects in process engineering,[8] while modern feature-selection methods provide principled ways to reduce redundant variables before modeling.[9] Ensemble learners such as RFs,[10] together with instance-based methods such as KNN,[11] are frequently adopted because they handle nonlinear relationships with limited assumptions; additionally, post-hoc explanation methods (e.g., shapley additive explanations (SHAP)) are increasingly used to interpret feature contributions and improve trust in ML models.[12]

These works collectively indicate that combining statistical feature selection methods with ML algorithms can provide meaningful insights and enable more accurate prediction models. Motivated by these findings, we evaluated the integration of ANOVA, PCA, KNN, and RF for optimizing thin-film quality in CVD processes.

3. Materials and Methods

ML techniques were applied in this study to reduce production yield losses and enhance testing efficiency. To ensure data quality, abnormal samples were removed through outlier

detection and data consistency checks before model training. The remaining process variables were standardized and subsequently used for feature selection and model construction. Figure 1 presents the overall workflow of the proposed method.

3.1 Data collection

Figure 2 outlines the complete seven-step CVD coating workflow adopted in this study. As shown in the process recipe, the procedure begins with Chamber Stabilization, followed by Chamber Cleaning and Condition, which together represent the preparatory stages before film deposition. The workflow then proceeds to the core coating stages, namely, Film Dep. (Growth) and Film Dep. (R), where the main deposition process is carried out in accordance with the experimental recipe. After deposition, the process enters the final thermal transition stages of Cooling and Cooling to RT, indicating the gradual return of the system to room-temperature conditions. Overall, this seven-step sequence provides a clear and structured representation of the experimental CVD process and serves as the procedural basis for the subsequent organization of the study.

As illustrated in Fig. 2, the CVD procedure was monitored throughout the full coating sequence. Signals captured by the process monitoring equipment were organized into four variable categories: time, pressure, temperature, and flow rate. Using the seven recipe stages

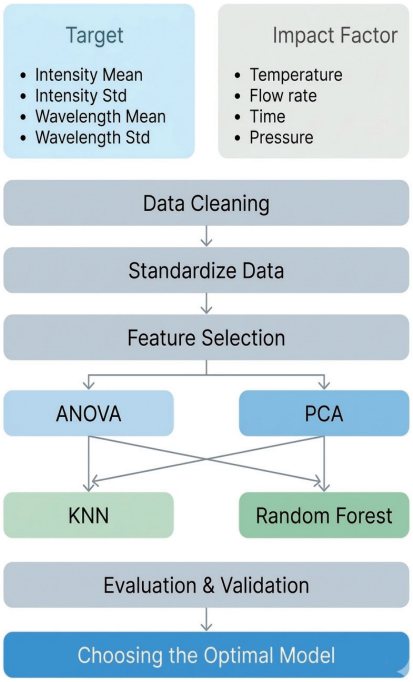

Fig. 1. (Color online) Process flowchart of model developed in this study.

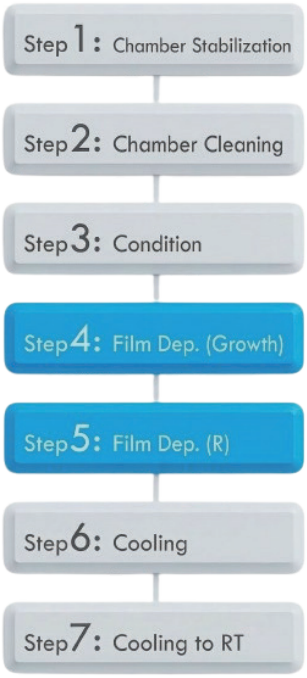

Fig. 2. (Color online) CVD coating process.

shown in Fig. 2, we established 13 parameter definitions, which together produced the 91 features summarized in Fig. 3.

Aside from the process-side variables, photoluminescence (PL) measurements describing the quality of WS_2 thin films were also collected, and production energy consumption was treated as an additional target. As shown in Fig. 4, the PL outputs can be grouped into four evaluation indices.

- Mean intensity – the average of the maximum PL intensity, representing luminescence performance.
- Intensity standard deviation (STD) – the uniformity of luminescence intensity.
- Mean wavelength – the average wavelength corresponding to the peak intensity, indicating emission characteristics.
- Wavelength STD – the uniformity of the emission wavelength distribution.

These four metrics serve as the primary prediction targets owing to their importance in assessing thin-film optical quality.

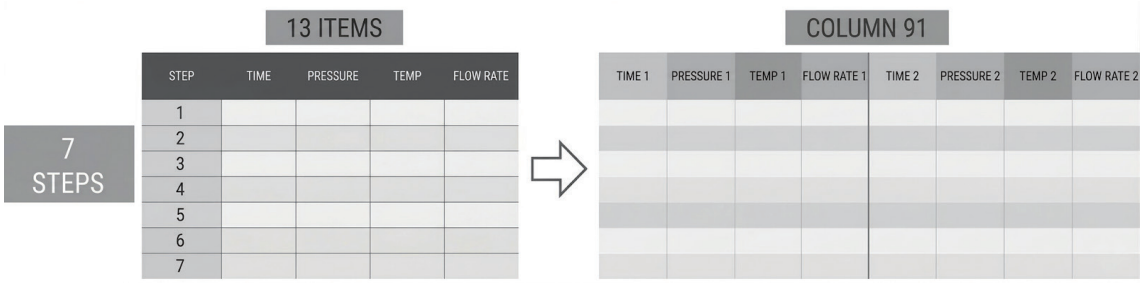

Fig. 3. Dataset constructed in this study.

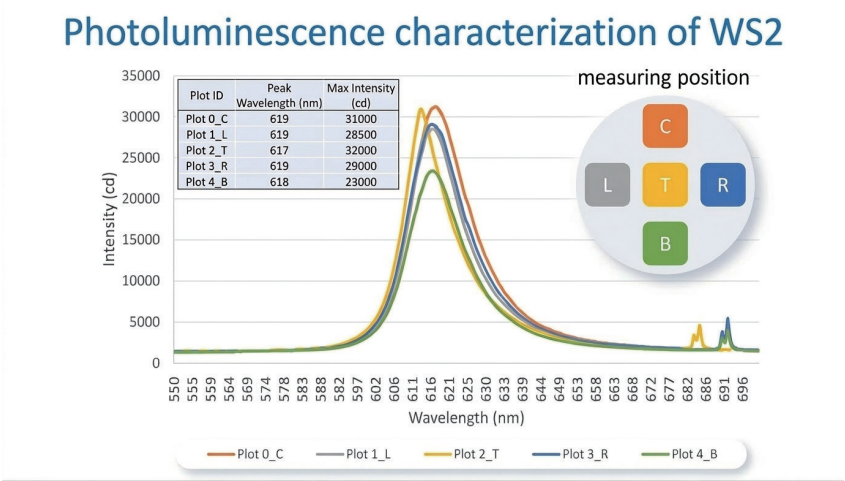

Fig. 4. (Color online) WS_2 thin-film quality data.

3.2 Data preprocessing

Several preprocessing operations were carried out before model development to improve data reliability. Missing numerical entries were filled with mean or median values in accordance with the distribution of each feature, and outlying samples were detected from the data distribution and removed to limit their influence on subsequent statistics and predictions.

Before feature selection and model fitting, all process variables were transformed to a standardized scale using

$$z = \frac{(x - \mu)}{s}, \tag{1}$$

where μ represents the mean of the training data, and s denotes STD. Equation (1) was used for feature centering and standardization by removing the mean and scaling to unit variance.[13]

Standardization is essential for ML algorithms because models such as ANOVA-based feature ranking, PCA, and distance-based methods such as KNN are sensitive to differences in feature scales. Without normalization, variables with larger numerical ranges dominate the optimization process and degrade model accuracy.

3.3 ANOVA

ANOVA was employed to examine whether the response differed significantly across feature groups.[14] In this study, the method uses the F statistic between each candidate parameter and the target variable to quantify how strongly that parameter is associated with the response. Because it offers a direct statistical view of factor relevance, ANOVA is suitable for screening influential settings in thin-film process data.

3.4 PCA

PCA was used to compress the information contained in the high-dimensional feature set into a smaller number of orthogonal components.[15] It does so by projecting the data onto directions that explain the largest amount of variance, thereby retaining the dominant structure of the dataset in a reduced space. This makes PCA useful for summarizing correlated process variables before model construction.

3.5 KNN

KNN is a supervised nonparametric algorithm applicable to both classification and regression. The method yields an estimate of the output of a query sample from nearby observations in the feature space: majority voting is used for classification, whereas neighboring target values are averaged for regression. Since the model stores the training data and performs computation mainly during inference, KNN is commonly described as an instance-based or lazy learner.[16]

In classification tasks, the predicted label of KNN is defined as

$$y_q = mode\, y_{i1}, y_{i2}, \ldots y_{ik}, \tag{2}$$

where y_q is the predicted label for the query point, and i_1, i_2, ..., i_k are the indices of the k nearest neighbors.

3.6 RF

RF is a supervised ensemble method for classification and regression that aggregates the outputs of multiple decision trees. Each tree is built from a subset of the data, and the final prediction is obtained by averaging their results, which generally improves robustness and generalization. Increasing the number of trees can enhance stability and accuracy, although it also raises computational demand.[17]

The output of the RF model is calculated as

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x), \tag{3}$$

where $\hat{y}$ is the predicted output for the input vector, T is the total number of decision trees, and $f_t(x)$ is the output of each decision tree.

3.7 R^2 score

The R^2 score measures how well the predicted values explain the variation in the observed data.[18] Its value ranges up to 1, and a value nearer to 1 indicates that the model provides a better fit to the true targets.

The R^2 score is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\hat{Y} - Y_i)^2}, \tag{4}$$

where X_i is the ith predicted value of the model, Y_i is the ith actual value, $\hat{Y}_i$ is the mean of actual values, and m is the number of samples.

3.8 Mean squared error (*MSE*)

MSE quantifies the average squared gap between model predictions and ground-truth values. Because larger errors receive a stronger penalty, this metric is useful for judging regression quality and identifying cases where prediction deviations are substantial [Eq. (5)].

MSE is defined as

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(X_i - Y_i\right)^2, \tag{5}$$

where X_i is the ith predicted value of the model, Y_i is the ith actual value, and m is the number of samples.

4. Results and Discussion

4.1 Feature selection

Figures 5–8 summarize the feature-ranking results for the four PL targets: mean intensity, intensity STD, mean wavelength, and wavelength STD. In each figure, the left panel shows the ANOVA F-score ranking, whereas the right panel presents the PCA-based ranking derived from principal-component contributions. These plots were used to identify a reduced set of influential process variables for later modeling and to compare whether each target depends on a small number of dominant factors or a broader combination of parameters.

4.2 Parameter importance

- Mean intensity
  The ANOVA results showed that the importance was in the order of temperature > flow rate > time > pressure.
  The PCA results showed that the importance was in the order of temperature > pressure > time > flow rate.
- STD of intensity
  The ANOVA results showed that the importance was in the order of flow rate > temperature > pressure = time.
  The PCA results showed that the importance was in the order of pressure > time > flow rate > temperature.

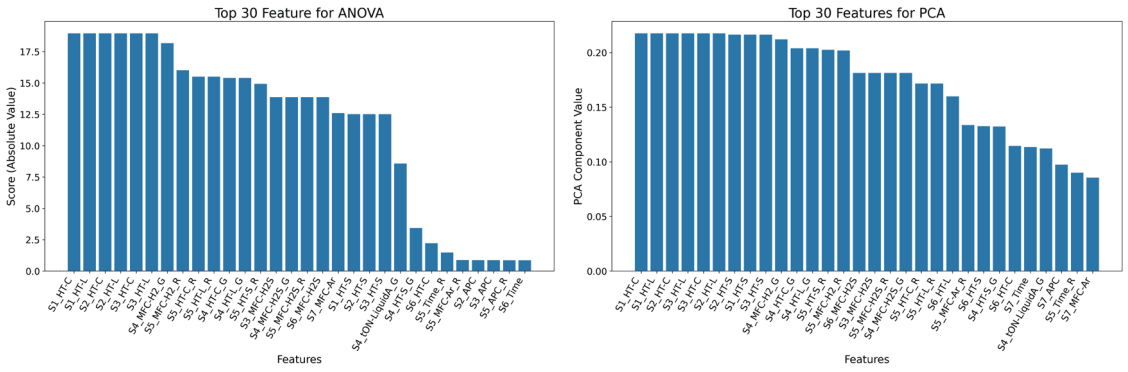

Fig. 5. (Color online) Feature selection using intensity mean and results of ANOVA and PCA.

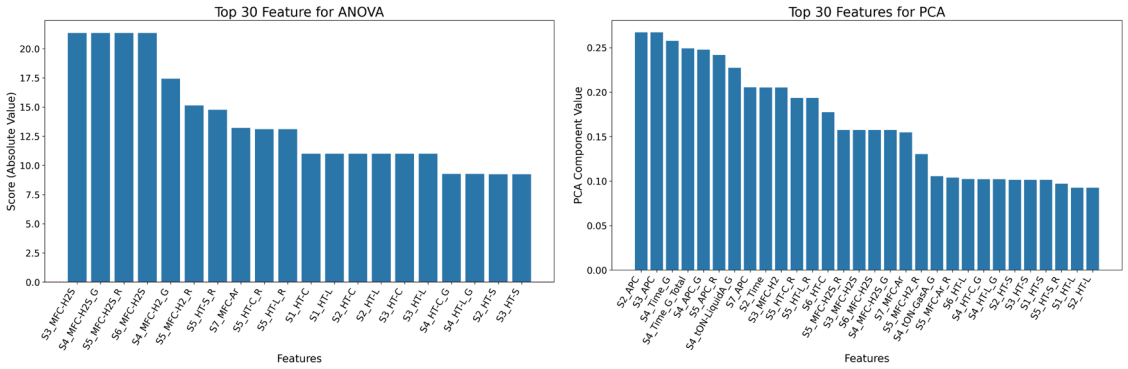

Fig. 6. (Color online) Feature selection using intensity STD and results of ANOVA and PCA.

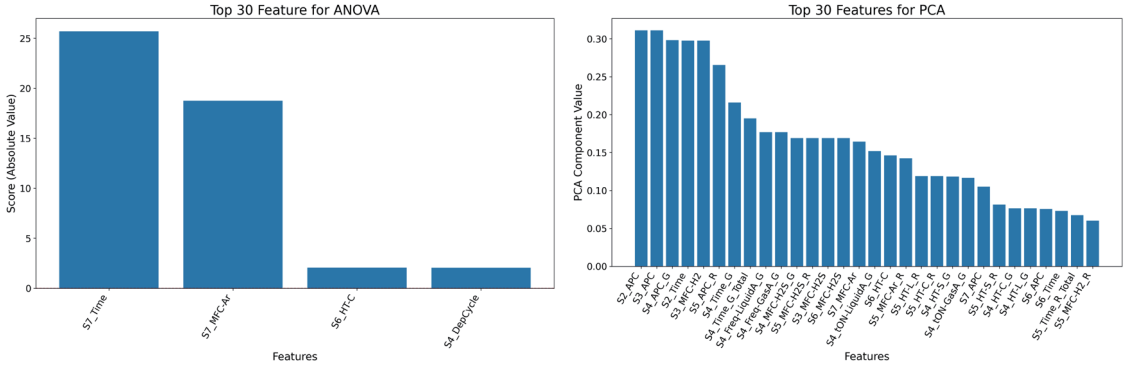

Fig. 7. (Color online) Feature selection using wavelength mean and results of ANOVA and PCA.

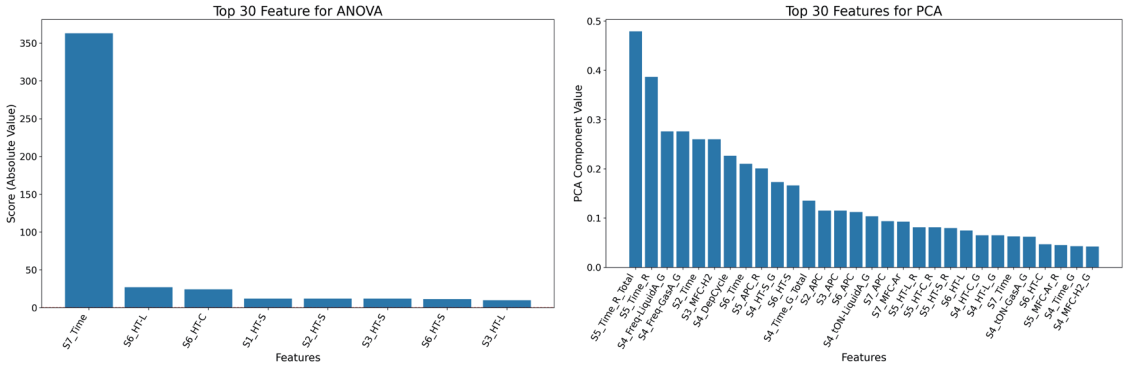

Fig. 8. (Color online) Feature selection using wavelength STD and results of ANOVA and PCA.

- Mean wavelength
  The ANOVA results showed that the importance was in the order of time > flow rate> temperature > pressure.
  The PCA results showed that the importance was in the order of pressure > time > flow rate > temperature.

- STD of wavelength
  The ANOVA results showed that the importance was in the order of time > temperature> pressure = flow rate.
  The PCA results showed that the importance was in the order of time > pressure > temperature > flow rate.
  Through the comparative analysis of parameter importance, influencing factors for each feature were identified as follows.
- Mean intensity
  Temperature influenced the mean intensity indicating that controlling the chamber temperature is crucial.
- STD of intensity
  Pressure played a key role in the STD of intensity despite the flow rate being highlighted in ANOVA.
- Mean wavelength intensity
  The PL wavelength was affected by the processing time.
- STD of wavelength
  The variation in wavelength was affected by the deposition time, according to both methods.

4.3 Selection result

The evaluation metrics of different algorithms were compared as follows. Table 1 presents the comparative performance results for the mean intensity. The RF-based models achieved substantially lower *MSE* and higher R^2 than KNN-based models. The RF + PCA model demonstrated excellent results with an *MSE* of 2,711,716.500 and an R^2 of 0.882. In contrast, the KNN-based models yielded lower R^2, with the KNN + PCA model at 0.510 and the KNN + ANOVA model at 0.451 but with higher *MSE*. Overall, RF-based models demonstrated superior predictive accuracy and error reduction compared with KNN-based models.

Table 2 shows the results for intensity STD. RF + PCA achieved the highest R^2 value at 0.872, whereas RF + ANOVA yielded the lowest *MSE* at 2,805,645.368 with R^2 = 0.858. The KNN variants showed weaker predictive capability overall, especially KNN + PCA, which recorded the lowest R^2 of 0.662 and the largest error. Taken together, the results again favor RF-based modeling for this response variable.

Table 3 presents the performance results for model combinations. The RF + ANOVA model showed the best performance, with the lowest *MSE* of 4.600 and the highest R^2 of 0.837. The RF + PCA model demonstrated excellent results, with an *MSE* of 7.185 and an R^2 of 0.753. The

Table 1
Evaluation metrics of mean intensity.

Model	*MSE*	R^2
KNN + PCA	82869564.716	0.510
KNN + ANOVA	31340994.863	0.451
RF + PCA	2711716.500	0.882
RF + ANOVA	7940668.455	0.936

Table 2
Evaluation metrics of STD intensity.

Model	*MSE*	R^2
KNN + PCA	6942605.172	0.662
KNN + ANOVA	513097.834	0.725
RF + PCA	3335175.686	0.872
RF + ANOVA	2805645.368	0.858

Table 3
Evaluation metrics of mean wavelength.

Model	*MSE*	R^2
KNN + PCA	21.195	0.537
KNN + ANOVA	14.862	0.687
RF + PCA	7.185	0.753
RF + ANOVA	4.600	0.837

KNN-based models showed lower performance, with the KNN + ANOVA model achieving an *MSE* of 14.862 and an R^2 of 0.687, while the KNN + PCA model yielded the lowest R^2 of 0.537 and the highest *MSE* of 21.195. The RF + ANOVA model outperformed other models in terms of predictive accuracy and error minimization.

Table 4 shows the performance results specifically for the STD of wavelength. The RF + PCA model presented the best performance, with the lowest *MSE* of 2.064 and the highest R^2 of 0.842. The RF + ANOVA model performed well, with an *MSE* of 2.411 and an R^2 of 0.837. In contrast, the KNN-based models yielded lower predictive performance. The KNN + ANOVA model showed an *MSE* of 19.942 and an R^2 of 0.290, whereas the KNN + PCA model presented the highest *MSE* of 23.021 and the lowest R^2 of 0.217. The RF + PCA model outperformed other models in terms of predictive accuracy and error minimization.

In this study, comparisons with previous studies highlight its novelty. Prior fab-oriented studies commonly treated film thickness estimation as a virtual metrology problem, leveraging rich sensor or plasma signals to predict thickness without offline measurements.[3–6] In contrast, in this study, we focused on a compact feature-engineering pipeline using routinely recorded CVD process parameters, combining standardization, PCA, and ANOVA-based factor screening to support fast quality monitoring when sensing is limited.[8,9] From the modeling perspective, our results are consistent with the broader observation that ensemble learners such as RFs are strong default baselines for nonlinear process data,[10] while instance-based methods such as KNN can be competitive but are more sensitive to feature scaling and neighborhood structure. [11] Recent semiconductor thin-film studies further emphasize the value of explainability (e.g., SHAP) on top of tree-based models to turn predictions into actionable process insights.[12,19] Compared with simulation-driven approaches—such as ML-assisted CALPHAD or thermodynamic surrogates that accelerate high-dimensional CVD analysis—our contribution is a practical, data-efficient framework validated on experimental coating data and explicitly tied to statistical factor screening.[20,21] Finally, optimization-oriented work (e.g., Bayesian optimization for WS_2 CVD growth) suggests a clear next step: extending the current predictive

Table 4
Evaluation metrics of STD of wavelength.

Model	*MSE*	R^2
KNN + PCA	23.021	0.217
KNN + ANOVA	19.942	0.290
RF + PCA	2.064	0.842
RF + ANOVA	2.411	0.837

baseline toward closed-loop recipe tuning once online metrology signals are available.[22] Overall, across the four PL quality targets, the RF-based models achieved R^2 values up to 0.936 (mean intensity), 0.872 (intensity STD), 0.837 (mean wavelength), and 0.842 (wavelength STD), and reduced MSE by 59.6 to 96.7% relative to the KNN + PCA baselines when comparing the lowest-MSE configurations for each target.

5. Conclusions

We explored the effectiveness of integrating ML for CVD process optimization in semiconductor manufacturing.[23] Owing to feature selection and modeling, the RF model combined with ANOVA or PCA outperformed the KNN model, especially in predicting PL-related metrics. The proposed data-driven method enhances process stability and yield and provides important data for intelligent, automated production lines adaptable to the manufacturing process using various materials.

Acknowledgments

This research was supported by projects under Grants No. NSTC 113-2221-E-032-030 and NSTC 113-2224-E-492-001.

References

1 F. T. Z. Toma, M. S. Rahman, K. M. A. Hussain, and S. Ahmed: J. Mod. Nanotechnol. **4** (2024) 100015.
2 Y. Fang, Y. Hu, X. Ma, and Y. Lu: J. Spectrosc. Explor. **32** (2021) 81. https://doi.org/10.23919/JSEE.2021.000008
3 M. H. Hung, T. H. Lin, F. T. Cheng, and R. C. Lin: IEEE/ASME Trans. Mechatronics **12** (2007) 308. https://doi.org/10.1109/TMECH.2007.897275
4 Y.-Y. Su, T.-H. Lin, F.-T. Cheng, and P.-J. Wu: IEEE Trans. Semicond. Manuf. **21** (2008) 426. https://doi.org/10.1109/TSM.2007.913222
5 H.-J. Roh, S. Ryu, Y. Jang, N.-K. Kim, Y. Jin, S. Park, and G.-H. Kim: IEEE Trans. Semicond. Manuf. **31** (2018) 232. https://doi.org/10.1109/TSM.2018.2824314
6 V. Maitra, Y. Su, and J. Shi: Expert Syst. Appl. **249** (2024) 123559. https://doi.org/10.1016/j.eswa.2024.123559
7 T. Lee and C. O. Kim: IEEE Trans. Semicond. Manuf. **28** (2015) 80. https://doi.org/10.1109/TSM.2014.2378796
8 D. C. Montgomery: Design and Analysis of Experiments (John Wiley & Sons, Hoboken, 2017) 9th ed., Chap. 8.
9 I. Guyon and A. Elisseeff: J. Mach. Learn. Res. **3** (2003) 1157. https://jmlr.org/papers/v3/guyon03a.html
10 L. Breiman: Mach. Learn. **45** (2001) 5. https://doi.org/10.1023/A:1010933404324
11 T. M. Cover and P. E. Hart: IEEE Trans. Inf. Theory **13** (1967) 21. https://doi.org/10.1109/TIT.1967.1053964
12 S. M. Lundberg and S.-I. Lee: Proc. Adv. Neural Inf. Process. Syst. **30** (2017) 4765. https://arxiv.org/abs/1705.07874
13 C. M. Bishop: Pattern Recognition and Machine Learning (Springer, New York, 2006) pp. 124–127.
14 N. Nurrahma and R. Yusuf: Proc. Int. Conf. Interactive Digit. Media (ICIDM 2020) (2020) 1–6. https://doi.

org/10.1109/ICIDM51048.2020.9339676
15 I. T. Jolliffe: Principal Component Analysis, 2nd ed. (Springer, New York, 2002) Chap. 6.
16 IBM: K-Nearest Neighbors Algorithm. https://www.ibm.com/topics/knn (accessed Feb. 2026)
17 J. K. Jaiswal and R. Samikannu: Proc. World Congr. Comput. Commun. Technol. (WCCCT 2017) (2017) 65. https://doi.org/10.1109/WCCCT.2016.25
18 V. Amaresh, H. N. Chethan, and M. S. Nagegowda: Proc. Int. Conf. Smart Technol. Syst. Next Gen. Comput. (ICSTSN 2022) (2022) 1.
19 Y. Shi, Y. Cai, S. Lou, and Y. Chen: Appl. Intell. **54** (2024) 246. https://doi.org/10.1007/s10489-023-05121-2
20 J. Wang, B. Xu, K. Lee, W. Huang, H. Wang, J. Peng, and M. Xu: Calphad **88** (2025) 102806. https://doi.org/10.1016/j.calphad.2025.102806
21 B. Xu, W. Huang, J. Wang, S. Zhang, Z. Xu, R. Tu, W. Li, J. Peng, and C. Wang: J. Cryst. Growth **637–638** (2024) 127727. https://doi.org/10.1016/j.jcrysgro.2024.127727
22 F. Zhang, R. Tamura, F. Zeng, D. Kozawa, and R. Kitaura: ACS Appl. Mater. Interfaces **16** (2024) 59109. https://doi.org/10.1021/acsami.4c15275
23 S. H. Liao, C. W. Chen, C. Y. Lin, J. H. Wang, and W. J. Chen: Int. J. Adv. Manuf. Technol. In Press. https://doi.org/10.1007/s00170-026-17554-3