IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 March 2025, accepted 21 April 2025, date of publication 23 April 2025, date of current version 9 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3563761

RESEARCH ARTICLE

Yield Diagnosis and Tuning for Emerging Semiconductors During Research Stage

CHUNSHAN WANG[1], ZIZHAO MA[1], YUXUAN ZHU[1], CHENSHENG JIN[1], DONGYU CHEN[1], CHUXIN ZHANG[1], YINING CHEN[2], (Associate Member, IEEE), WENZHONG BAO[1], AND YUFENG XIE[1], (Member, IEEE)

[1]State Key Laboratory of Integrated Chips and Systems, Fudan University, Shanghai 200437, China
[2]School of Integrated Circuits, Zhejiang University, Hangzhou, Zhejiang 311200, China

Corresponding authors: Yining Chen (yining.chen@zju.edu.cn), Wenzhong Bao (baowz@fudan.edu.cn), and Yufeng Xie (xieyf@fudan.edu.cn)

This work was supported by NSFC under Grant 61874031.

ABSTRACT The process of taking a new semiconductor device from the lab to the factory involves a lot of time, funds and manpower, a large portion of which is spent on device yield improvement. In recent years new methods have been tried to rapidly improve yields and using machine learning (ML) algorithms is one option. However, they usually require a large dataset, which is often unavailable at the device research stage, emerging semiconductors (e.g., 2D materials) are extremely costly to pilot. In this paper, we propose a yield diagnosis and tuning scheme based on ensemble learning and Bayesian optimization, which demonstrate outstanding performance even with a limited data volume. We use real 2-D semiconductor device fabrication process data for scheme evaluation. Experimental results show that the algorithm for yield prediction has achieved regression fitting results whose mean absolute error (MAE) is no more than 8 points and explained variance (EVAR) is no less than 0.62, this indicates that the model fits well on this dataset. We also remanufactured a batch of devices based on the yield tuning recommendations to validate the effectiveness of our approach. The test results indicated a final yield score of 86 points, after evaluating several key indicators such as mobility and hysteresis, resulting in a 62% improvement.

INDEX TERMS Semiconductor manufacturing, yield diagnosis, yield prediction, yield tuning, Bayesian optimization.

I. INTRODUCTION

With the continuous improvement of semiconductor process technology, it has now entered the process node of about 5nm, and the industry is undergoing product yield problems in both ramp-up and maturation phase. The devices with low yields can hardly be applied in large scale and remain in the laboratory research stage [1], [2], [3], [4], [5], [6]. Thus, manufacturing companies and research institutes are looking to achieve high yields as soon as possible to enhance their competitiveness. Especially in the research phase of devices with new structures and materials according to Moore's Law, improving yield is the top priority.

To better illustrate this issue, we will give a brief overview of the manufacturing process of semiconductor devices. The process can be divided into multiple steps (assume N steps). Among these steps, some may be a single choice of multiple options, while others may be physical variables that can be continuously adjusted. The devices will be evaluated by measuring their electric performance after manufacturing. Then those raw device data will be scored by factory engineers. Fig. 1 illustrates the whole process.

Therefore, researchers are counting on yield diagnosis and yield tuning [7], [8], [9], [10], [11], [12] to rapidly improve the efficiency of semiconductor manufacturing. Yield diagnosis is divided into two sub-items. First is the root cause analysis, that is, analyzing which step or steps in semiconductor manufacturing have the greatest impact on the yield.

The associate editor coordinating the review of this manuscript and approving it for publication was Jiajie Fan.

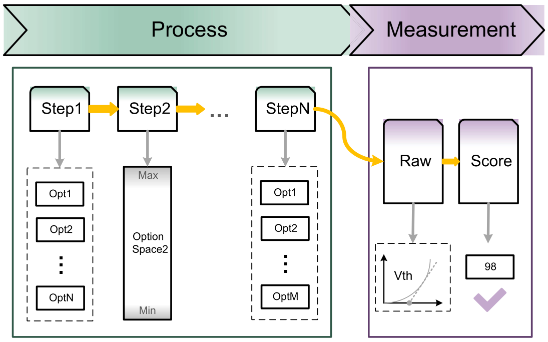


FIGURE 1. Semiconductor process and measurement steps.

The second is yield prediction. Given the steps involved in semiconductor manufacturing and the parameter selection at each stage, a vast number of process combinations can be generated, allowing us to predict the yield for each configuration. If researchers know which steps are important, they can adjust them in a more proactive manner. In addition, with yield prediction, workers can know the predicted yield of an adjusted manufacturing process in advance without the need for experimentation, which saves labor and material resources. Moreover, yield tuning directly gives you how to adjust the process step given by the algorithm to improve yield. The combination of yield diagnosis and yield tuning is the ultimate solution for improving semiconductor manufacturing yields.

In the past, yield diagnosis and yield tuning relied heavily on the experience of device experts. This stage required a lot of time and effort to repeat the experiments to obtain the optimal process combination for maximizing yield. To shorten this cycle as well as reduce the amount of labor invested, researchers are exploring new approaches and machine learning is one of them. The increasing sophistication of machine learning has enabled many previously intractable and complex problems to be well modeled and to make predictions valid enough to guide real-world engineering practices, thus facilitating scientific discovery and technological innovation like never before. Some researchers have adopted machine learning algorithms for Yield Diagnosis and Yield Tuning [13], [14], [15], and all have made some progress.

In paper [16], a novel statistical approach was proposed to address the yield improvement problem, taking into account the complexity of wafer fabrication systems. This method allows for the testing and analysis of data from a single defect source. Paper [17] introduced a statistical fault diagnosis technique that utilizes only the first or previous fault test vectors. This method accurately correlates the electrical fault location with inline inspection data, thereby enhancing yield efficiency. Paper [18] presented a factorial parameter screening algorithm aimed at improving manufacturing yields by adjusting key parameters that significantly affect yield. This approach focuses on identifying critical parameters in the manufacturing process of disk drives. The impact of process variations on yield improvement in active circuit design was explored in paper [19]. The authors proposed load-traction design methods that optimize load impedance to achieve high yield while maintaining strong output performance.

The existing machine learning algorithms used in yield diagnosis research include simple regression method [20], Random Forest [21], Naive Bayesian Classifier (NBC) [22], neural network [23] and process-aware DTCO [24]. Paper [20] uses a regularized regression model to simplify training parameters for yield prediction in multi-step manufacturing. However, it is only tested in a virtual process with three steps, making its results less convincing, particularly for advanced semiconductor manufacturing. Additionally, the prediction accuracy decreases as the number of predicted data points increases. Paper [21] uses LASSO regression to filter key operations and Random Forest (RF) for ranking Fault Detection and Classification (FDC) parameters. In their experiment, this method achieved 97.85% accuracy on a dataset with 325 wafers and nearly 8000 operations. Paper [22] proposed a new Bayesian network structure to predict product yield and detect faults. However, this method overlooks the impact of data collection and data loss in real process conditions. Paper [23] presents a Convolutional Neural Network (CNN) model, FDC-CNN, for fault detection. It links the output of the first convolutional layer to the structural meaning of the training data to locate faults. However, this method is applied to single-process semiconductor fabrication and requires a large amount of training data for effective network training. Paper [24] explores WS2-based 2D devices with a process-aware DTCO approach that improves inverter performance by ∼40% over silicon devices at IMEC 2nm nodes Reviewing the methods proposed previously, almost all of them demand a large data set and high data integrity, which are not suitable for semiconductor yield diagnosis and yield tuning in the exploration stage, as the volume of the new device data-set is generally too small to satisfy the demand.

TABLE 1. Comparison of common algorithms for yield diagnosis.

Algorithms	Small dataset adaptation	Interpret ability	Nonlinear Relationship Processing	Training speed
Simple regression	Medium	High	Bad	Fast
NBC	Low	Medium	Bad	Fast
Neural Network	Too Low	Low	Good	Slow
Process-aware DTCO	Medium	High	Medium	Slow
XGBoost/RF	High	Medium	Good	Fast

In the field of semiconductor industry, the evaluation of each parameter necessitates the utilisation of an actual tape out, accompanied by a significant experimental cost. Consequently, the selection of algorithms must be made with the objective of approximating the optimal solution with a minimal number of experiments, thereby facilitating the pursuit of the lowest cost [25]. The Bayesian optimization method utilises the feedback information from each experiment during the iteration process, employing black-box modelling of the objective function through the construction of an agent model, such as a Gaussian process (GP), thereby efficiently balancing exploration and utilisation on a global scale. The utilisation of a Gaussian process model for uncertainty representation is effective in capturing the variability present in process parameters and yield, attributable to fluctuations such as the chaotic effect observed in plasma etching [26]. The ability of Bayesian optimization to leverage the black-box fitting capability of its agent models is particularly advantageous in scenarios where data is limited and experimental costs are high, as evidenced by the following table, which provides a comparison of common machine learning algorithms for yield tuning:

TABLE 2. Comparison of common algorithms for yield tuning.

Algorithms	Sample efficiency	Noise Resistance Capacity	High-dimensional adaptability	global convergence
Bayesian optimization	High	Strong	Low	Strong
Grid Search	Low	Weak	Low	No Guarantee
Random Search	Low	Medium	Medium	No Guarantee
Genetic Algorithm	Medium	Medium	High	Medium

Grid search optimization traverses all parameter combinations, resulting in the consumption of a large number of computational resources, and random search is usually unable to accurately and quickly find the optimal parameter combinations due to its purely random nature; therefore, grid search and random search are not only inefficient in terms of sample efficiency, but also lack the theoretical guarantee of global convergence, while Bayesian optimization has a strong global search capability in theory and can avoid falling into local optimum [27]; genetic algorithms, although they perform better in dealing with high-dimensional data, are still not as fast in terms of convergence speed and sample efficiency as Bayesian optimization. Therefore, in semiconductor manufacturing, a scenario where each evaluation is expensive and the environment is complex, Bayesian optimization becomes a more appropriate choice.

For the research stage, to overcome challenges such as small amounts of data and unconsolidated parameter combinations, we propose a new scheme for yield diagnosis and yield tuning. In this proposed scheme, we apply the decision-tree based ensemble learning methods such as XGBoost and Random Forest for yield diagnosis in the device research phase as well as yield ramp-up phase. In addition, we use the Bayesian optimization method for yield tuning and real 2D semiconductor manufacturing process data at the laboratory research stage to test our algorithm. Experimental results indicate that the Mean Absolute Error (MAE) remains below 8%, while the Explained Variance (EVAR) exceeds 0.62 in the yield prediction test. The yield tuning experiment is tested on actual manufactured devices, which test results showed a final yield score of 86 points, achieving a 62% yield improvement.

II. PROPOSED SCHEME FOR YIELD DIAGNOSIS AND YIELD TUNING

The overall framework of the proposed scheme is shown in Fig.2. It primarily comprises two components: the Yield Diagnosis Model and the Yield Tuning Model. Initially, we will use the parameters of each step in the device manufacturing process and the electrical characteristics measured as training data sets to train the Yield Diagnosis Model. During the training process, we optimize the model based on the results of the yield prediction given by the model until the model is able to fit the dataset well. Then we can apply this model to find the root cause that affects the final yield. After that, the Bayesian optimizer can be used to find the best yield among the various process combinations and feed that combination into the Yield Diagnosis Model for yield prediction to determine if the yield is improved or not. This process usually requires iterations, and when the maximum number of iterations is reached, the final process combination is the optimal one after Yield Tuning.

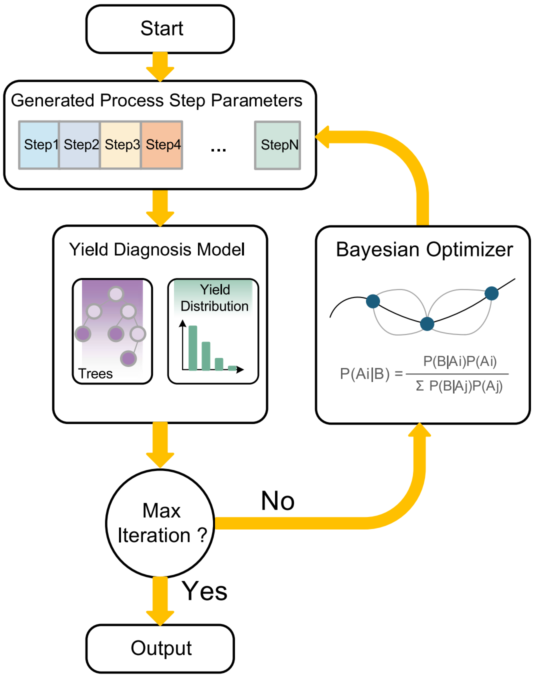


FIGURE 2. Framework of proposed yield diagnosis and tuning scheme.

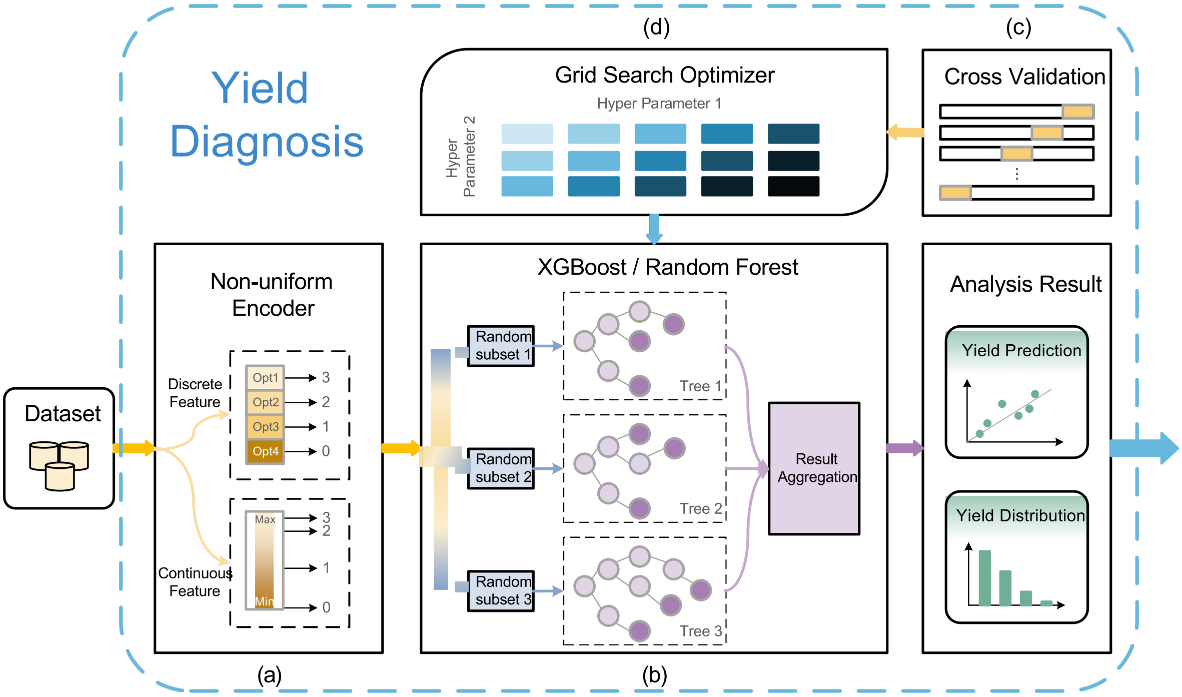


FIGURE 3. Yield Diagnosis main flow. (a) Non-uniform encoder; (b) Ensemble Learning model; (c) Cross Validation;(d) Grid Search optimize.

The following parts of this section provide a detailed description of the algorithmic principles involved in these two models.

A. YIELD DIAGNOSIS MODEL

Fig.3 illustrates the details of the Yield Diagnosis Model. It mainly consists of four parts, which are Non-uniform Encoder(a), Ensemble Learning Model(b), Cross Validation(c) and Grid Search Optimizer(d).

1) NON-UNIFORM ENCODER

Although there are continuous adjustable parameters in semiconductor manufacturing, it is impossible to try a large number of choices during manufacturing in the device research stage and only a few sets of parameter options are tried. Therefore, it is actually discrete data and we should encode them first. It is easy to understand the encoding of the discrete option parameter itself. The encoding of continuous parameters is actually a space transformation operation. Space transformation is common in Euclidean space-based machine learning algorithms, such as Support Vector Machines (SVMs). SVMs often use a Kernel Space with several options, such as the Radial Basis Function (RBF) and Polynomial kernels. They are all space transformations. For decision tree-based algorithms, the encoding of continuous parameters has little effect. But encoding has a large impact on SVM, KNN and other Euclidean space-based algorithms.

Our algorithm has a non-linear encoder for continuous parameters. We do not divide the continuous space equally (like quantization); instead, we sort the value from small to large and encode the sorted values from zero to the count minus one.

2) ENSEMBLE LEARNING MODEL

The idea of ensemble learning is that single machine learner cannot achieve optimal results but multiple learners can work together to overcome the shortcomings of a single learner. Ensemble Learning has been demonstrated to efficiently fuse the experiences of multiple small samples, thereby enhancing the model's generalization capability. This is achieved by artificially introducing random parameters during the early development stage of the process and incorporating the experimental results into the learning process. The overreliance on extensive training data sets can be circumvented by leveraging small-sample learning methodologies, which facilitate the identification of local optima through a limited number of random experiments. This approach serves to effectively mitigate the overfitting problem. Furthermore, in the late stages of semiconductor process optimization, the adjustment of process parameters is typically minimal, ensuring that even in the event of overfitting, the outcomes align with the criteria for real-world production decisions. We use cross-validation and regularization techniques, feature selection and dimensionality reduction, and hyperparameter tuning to further reduce the risk of overfitting.

Taking the decision tree-based ensemble learning algorithm as an example, firstly, divide the data set into random subsets. Use these random subsets to train different decision tree learners. The learning results of each decision tree learner are different. An aggregation algorithm is required to integrate the training results of each learner to obtain the final result. The aggregation algorithm is generally weighted sum of prediction results of each sub-decision tree.

The ensemble learning expression is shown below.

$$\hat{y}_i = \sum\nolimits_{K-1}^{K} f_k\left(x_i\right) \tag{1}$$

Mature methods for training decision trees include algorithms such as ID3, C4.5, and CART. There are many algorithms for training ensemble learners such as XGBoost and Random Forest. We use ensemble learning to overcome the problem of low accuracy of a single decision tree and use XGBoost/Random Forest to further improve the accuracy of prediction which can also simplify the complexity of each sub-decision. Although XGBoost and Random Forest do require a certain amount of computation time during the training process, their online inference is fast and fully meets the manufacturing industry's need for fast response. Once the training of the Ensemble Learning model is complete, yield prediction and yield distribution can be performed based on the model.

3) MODEL OPTIMIZATION METHOD FOR YIELD PREDICTION AND DISTRIBUTION

As shown in Fig. 4, we first perform non-uniform coding after obtaining the raw data. Then we use the Cross-Validation method to generate the training data set and the evaluation data set from the encoded data set, and use the Grid Search method to generate the hyperparameters needed for the machine learning algorithm. Then XGBoost regressor is applied to load these generated hyperparameters, and then train the regressor using training data set to generate a model.

We use this model to predict the evaluation data set, and then compare it with the actual results to calculate MAE and EVAR, the corresponding equations are as follows.

$$\mathrm{MAE}\left(y, \hat{y}\right) = \frac{1}{n_{\mathrm{samples}}} \sum\nolimits_{i=0}^{n_{\mathrm{samples}}-1} \left|y_i - \hat{y}_i\right|. \tag{2}$$

$$\mathrm{EVAR}\left(y, \hat{y}\right) = 1 - \frac{\mathrm{Var}\{y - \hat{y}\}}{\mathrm{Var}\{y\}} \tag{3}$$

These two indicators reflect the explanatory effect of the model, and a lower MAE and a higher EVAR both indicate a better fit of the regression model on the given data set. The Pseudocode of Yield Diagnosis modeling algorithm is shown below in Fig. 5.

This is just one selection of hyperparameters and one training result of data set partitioning. Using grid search and cross-validation we can get all combinations of parameters and all divisions of the training set, and thereby we can obtain training results in all cases. We compare all MAEs and select the training model with smallest MAE as the final optimized model, and the yield prediction based on this model can be considered as the most accurate.

Besides, with this model we can get the yield distribution of semiconductor manufacturing. From the yield distribution we can clearly see the effect of a process step with different parameters on the final yield. Thus, Yield Diagnosis can be performed on the basis of this model.

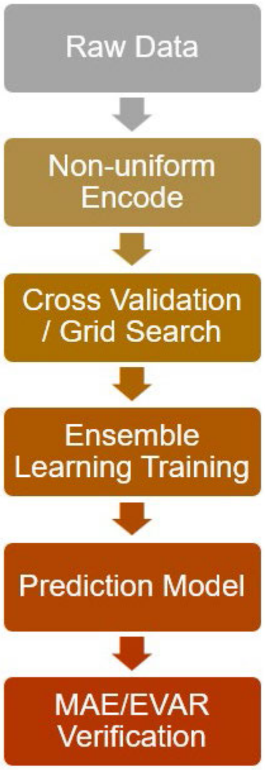


FIGURE 4. Flow chart for yield prediction.

```
Algorithm 1 Yield Diagnosis Modeling Algorithm
Input :     D : the manufacturing yield dataset
Output :    m*: the machine learning fitting model
Set model algorithm to XGBoost Regressor
Initialize grid search optimizer opt
Initialize cross validation spliter cv
Initialize model m
D_enc = NonUniformEncode(D)
repeat
    hyper_params = opt.Generate()←Generate model hyper parameters by optimizer;
    m.SetParams(hyper_params)←Set hyper parameters of model m;
    repeat
        D_train,D_eval=cv.split(D_enc)←Delineate the cv dataset;
        m.Train(D_train);
        mae = model.MAE(D_eval)←Calculate the model mae ;
    until cv traverses D_enc
until opt generates all the combinations of hyper parameters;
Choose the best model m* with the lowest mae;
return m* with Yield Prediction and Yield Distribution;
```

FIGURE 5. Yield diagnosis modeling algorithm.

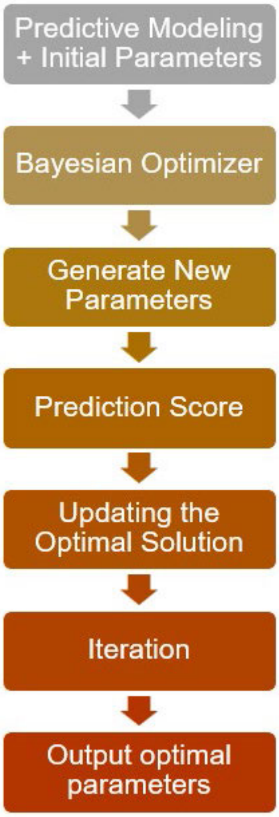

Predictive Modeling + Initial Parameters
Bayesian Optimizer
Generate New Parameters
Prediction Score
Updating the Optimal Solution
Iteration
Output optimal parameters

FIGURE 6. Flow chart for yield tuning.

```
Algorithm 2 Yield Tuning Algorithm
Input:   $m^*$: the yield diagnosis model; D: the semi-conductor
         manufacturing yield dataset; $n$: number of max
         iterations; the manufacturing yield dataset;
Output:  collections of recommended manufacturing
         parameters ;
Initialize a Bayesian Optimizer $opt$
Initialize an empty collection $P_{rec}$
Choose one set of parameters $p_{best}$ from $D$ with the highest
score $s_{best}$
Guide the $opt$ to current best parameters set $p_{best}$
for $i$=1 to $n$ do
    $p_i$=$opt$.Generate()
    $s_i$=$m^*$.Predict($p_i$)
    if $s_i$ > $s_{best}$ then
        $P_{rec}$.Add($p_i$);
    end if
end for
return $P_{rec}$;
```

FIGURE 7. Yield tuning algorithm.

B. YIELD TUNING MODEL

The already trained yield prediction model contains yield distribution. Based on this, the parameter options of the process steps can be adjusted and combined into new process step parameter options, and then the model can be applied to perform yield prediction on these new combinations. If the predicted score is improved, it can be considered as a discovery of a better set of parameters.

First of all, in semiconductor manufacturing, the process steps are limited, and secondly, the options of each step parameter in the field of device development are discrete and limited. Therefore, using the Grid Search algorithm of brute force search to traverse each combination can theoretically find all combinations with yield improvements. However, once the process steps increase and the options for each step increase, the price paid by Grid Search will increase significantly.

We use a Bayesian optimizer, a heuristic approach, to efficiently identify the optimal parameter combination. Starting with the parameter set that achieves the highest initial score, the optimizer iteratively generates new combinations. It employs a yield diagnosis model to predict scores for these combinations. If a predicted score surpasses the current best, the optimizer identifies it as a superior parameter set. This process repeats, refining multiple iterations to discover better combinations.

Furthermore, Bayesian Optimisation functions as an asynchronous task in the back office, generating a daily batch of recommendations for optimisation parameters rather than directly interfering with real-time decision-making on the production line. During line changeovers or maintenance windows, process engineers can implement parameter updates in conjunction with equipment status, thus avoiding disruption to normal production rhythms. The yield tuning algorithm is outlined below.

III. YIELD DIAGNOSIS EXPERIMENT

A. EVALUATION ENVIRONMENT

To better illustrate our proposed scheme and verify its effectiveness, we use real 2D process data for testing. The steps and parameters are shown in Fig. 8.

Give an example of real semiconductor manufacturing data. Totally, there are 19 steps in the real process including 47 parameters, of which 8 are discrete and 39 are continuous. The data set has 8 evaluation indicators, 4 of which are the original measurements of the physical measurement of the device data (raw data), the other 4 are score data after artificially scoring the 4 raw data. These 4 evaluation indicators are named hysteresis, μ, I_{on}/I_{off}, and Vth.

We use 47 F parameters to replace the actual process parameters, and some of the key parameters are shown in the table below:

Semiconductor devices usually have multiple optimization goals. Common goals are high-performance and low-power. High-performance indicators require high switching speed and high threshold voltage and better sub-threshold characteristics. Low-power goals require small leakage current and high requirement for the ratio of on-state current and off-state current. We also perform machine learning modeling on these two optimization indicators. Specifically, the total indicator of high-performance modeling comes up with 4 sub-items with ratio of 2: 2: 2: 4 of weighted sum which means threshold voltage is the most important factor, and the total indicator of

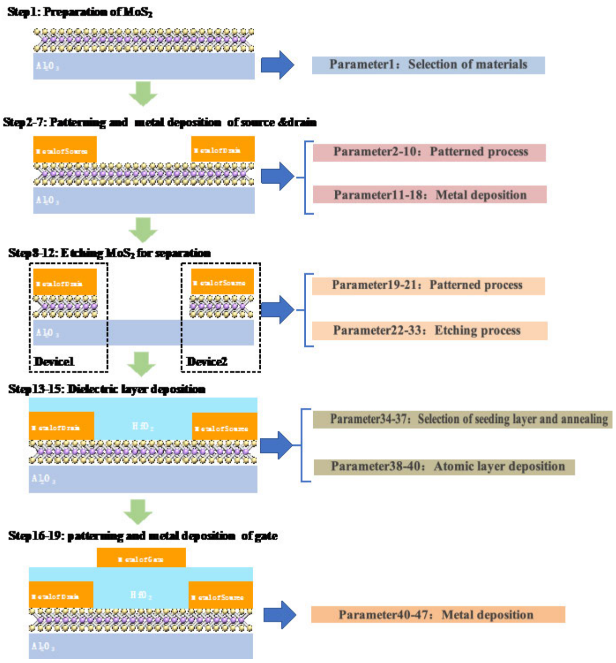

Step1: Preparation of MoS₂
Al₂O₃
Parameter1: Selection of materials
Step2-7: Patterning and metal deposition of source &drain
MetalofSource
MetalofDrain
Parameter2-10: Patterned process
Parameter11-18: Metal deposition
Step8-12: Etching MoS₂ for separation
Device1
Device2
Parameter19-21: Patterned process
Parameter22-33: Etching process
Step13-15: Dielectric layer deposition
HfO₂
Parameter34-37: Selection of seeding layer and annealing
Parameter38-40: Atomic layer deposition
Step16-19: patterning and metal deposition of gate
MetalofGate
Parameter40-47: Metal deposition

FIGURE 8. The fabrication steps and parameters of top-gate 2D device.

TABLE 3. Comparison of several F-parameters with the actual process.

F Parameter	Manufacturing process
F0	2D material batch
F10	Development time
F11	Source/Drain metal deposition Vaporization method
F14	Source/Drain metal deposition rate of Au
F23	Etching time
F24	Etching power
F34	Seeding layer
F35	Annealing temperature
F36	Annealing Time
F39	ALD Thickness
F41	Contact metal choice
F47	Top gate annealing

the low power modeling comes up with 4 sub-items with ratio of 2: 2: 4: 2, which means on and off current ratio is the most important factor. We use the algorithm steps mentioned in the previous section to train this real dataset and show the results.

As can be seen from Fig.9, the importance of the parameters under different indicators varies, for example, F34 is the most influential for hysteresis, I_{on}/I_{off} and Vth, while F0 is the most influential for μ. The importance of each parameter also changes after the combined evaluation of all indicators. For both the high performance and the low power model, F34 has the greatest influence, which reaches 122 points and 39 points respectively. But the subsequent ranking of the parameter's importance differed. For example, F39, F35, F36 are more important for high performance, but for low power F0, F41 and F1 are more important. Based on the F Score ranking we can easily identify the process steps that have an impact on the yield and adjust the parameters accordingly.

B. YIELD PREDICTION

The previous section introduced the analysis results of Root Cause, but the credibility of this result has yet to be evaluated, which can be converted into the accuracy of the model prediction. If the model prediction is accurate, the above Root Cause will be more credible.

After the training of XGBoost regressor is completed, we use this trained model to predict the pre-divided evaluation data set and compare the predicted value with the real value to evaluate the accuracy and credibility of the model training, that is, to calculate MAE and EVAR. To better illustrate the model's effect, we also drew Fig.10 to show the variance and fitting result.

The real value is used as the horizontal coordinate, and the predicted value of the model is the scatter plot of the ordinate, which is colored in lavender. This is similar to the confusion matrix in the classification problem. The scatters of absolute errors are also plotted in light gray on the same graph, allowing a clear view of situations of absolute error. For an excellent fit model, the lavender scatter should be concentrated as close to the diagonal line as possible and the absolute error value should be as small as possible. It can be seen from the subgraph (a) and (b) that the regression of the total score is good. Although some regions are deviated from the diagonal line, the high score region is concentrated, indicating that the high score region is more accurate. The model trained in Yield Prediction is also used in the Yield Tuning algorithm. Yield Tuning aims at increasing the prediction score, so it uses the high-scoring region of the Yield Prediction model and the scattered low score region has little impact on Yield Tuning.

In subgraph (c) and (d), the horizontal coordinate indicates each sample, the two curves on the ordinate are respectively the real and predicted values, and the gray shading is the absolute error. From all the subgraphs, the real scores appear to be concentrated within the 50 to 80 range. The absolute error remains minimal, and the statistical results confirm that the MAE of the trained regression model does not exceed 8%, while the explained variance stays above 0.62.

To compare the effects of different algorithms, we do more than just conduct experiments with XGBoost. We also tested other common machine learning algorithms, such as Euclidean space-based KNN, Ridge, Lasso, SVR, ElasticNet, and Probability-based Naive Bayesian regression, Gaussian process and neural network-based MLP (in this research, a fully connected neural network with 3 layers of multiple inputs and single output), and other decision tree-based ensemble learning RF, LightGBM.

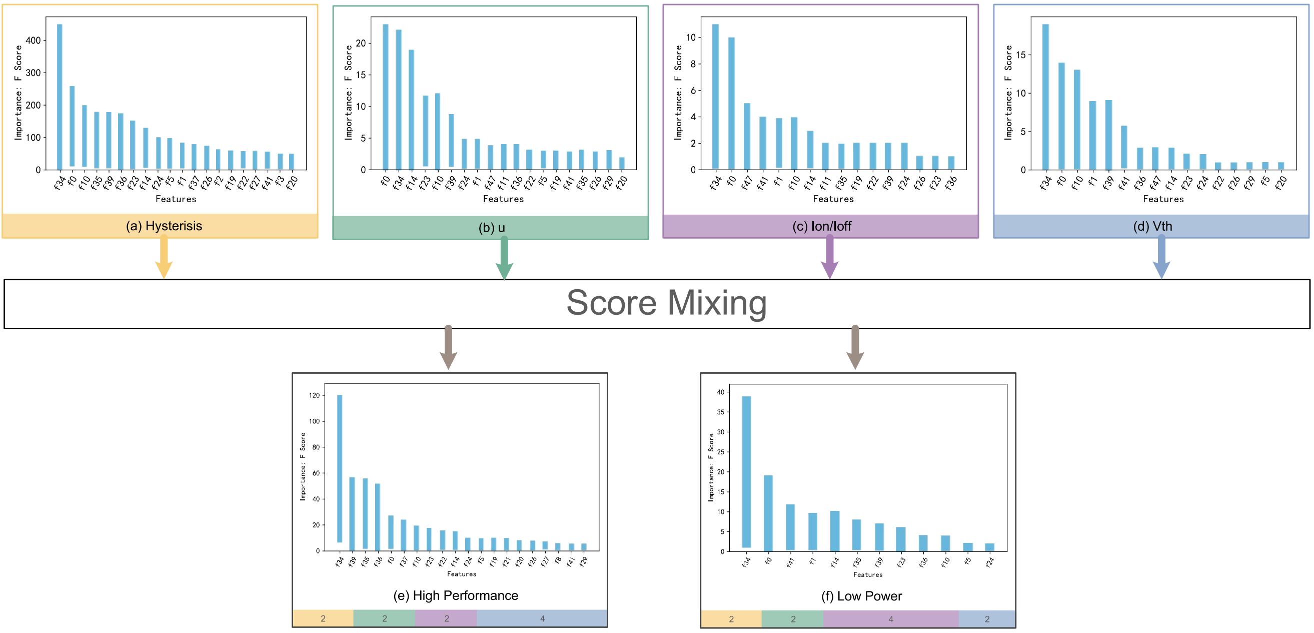
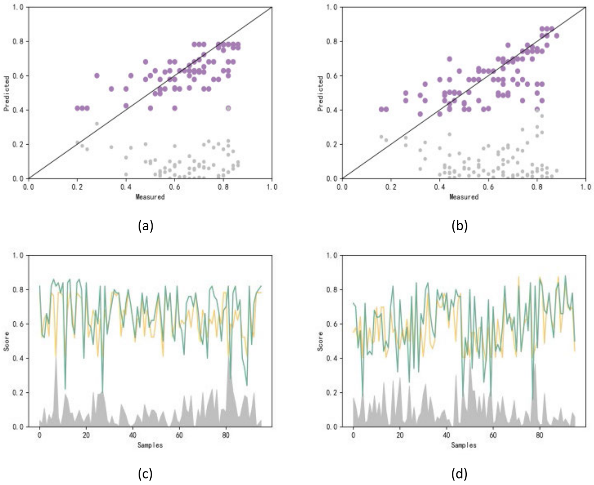


FIGURE 9. The F Score ranking in 4 indicators (a) hysteresis, (b) μ, (c) I_{on}/I_{off}, (d) Vth and mixed score of (e) High Performance benchmark and (f) Low Power benchmark.



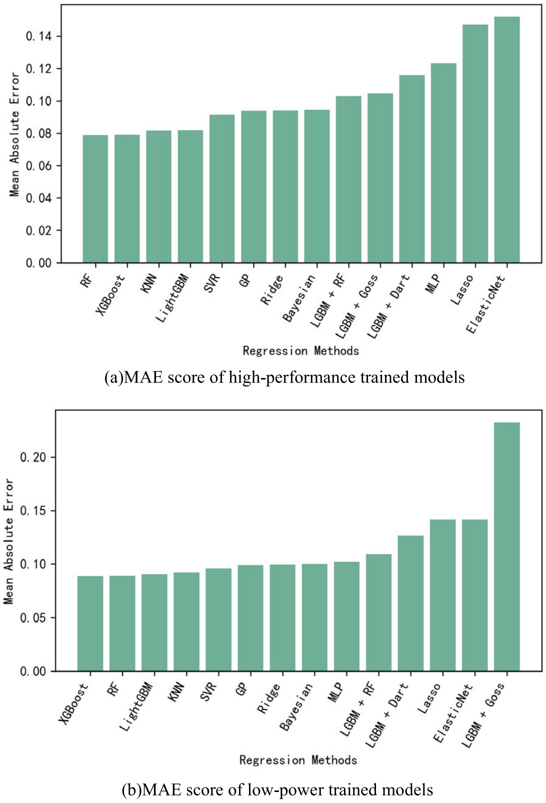
(a)

(b)

(c)

(d)

FIGURE 10. Model prediction result (a), (b) and fitting result (c), (d) under the Low Power and High Performance scoring benchmark respectively.



(a)MAE score of high-performance trained models

(b)MAE score of low-power trained models

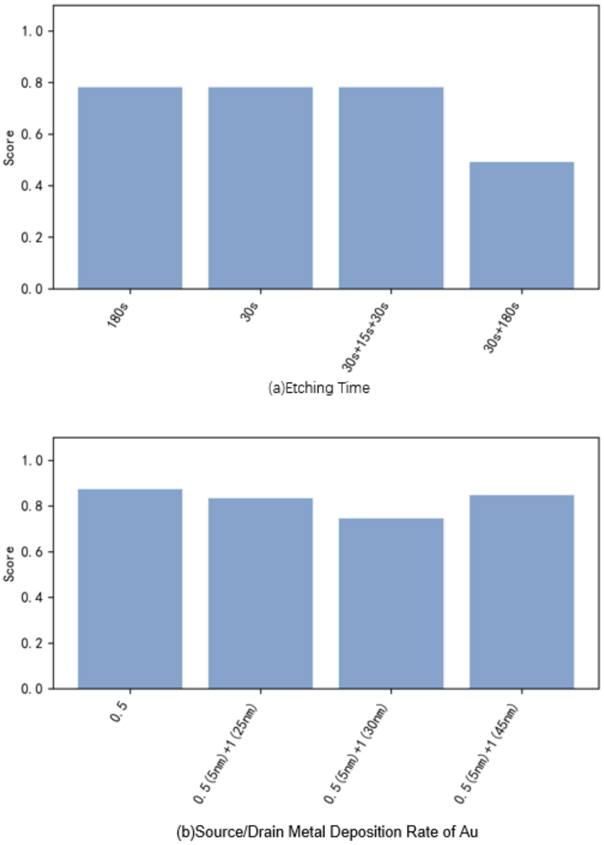
FIGURE 11. Different regression methods' MAE for (a) high-performance benchmark and (b) low-power benchmark.

We use MAE as an evaluation benchmark and the results are shown in Fig.11. (a) is the MAE score of models trained for high-performance optimization goal, while Fig.11. (b) presents the scores for low-power consumption optimization goal. The results indicate that RF and XGBoost achieve the best performance for both objectives, making it challenging to determine which is superior. This suggests that ensemble learning is particularly well-suited for semiconductor manufacturing at the research stage.

In the process of training MLP, there are cases of training failure. The reason is that the amount of training data is not enough. MLP is already a neural network with a small number of layers and scale, and the amount of training data is still insufficient, so larger networks such as AlexNet, ResNet are more difficult to train. Therefore, this fact further confirms that neural network algorithms are not suitable for the research phase yield diagnosis application scenario with low data volume.

C. YIELD DISTRIBUTION

With the model described above, we can predict the yield scores of the manufactured devices with different process parameter values, and thus derive the yield distribution. In this experiment, we only adjust the values for the key process parameters identified by root cause analysis, while keeping the default values for the other process steps. And thereby we can intuitively see the distribution of yield in several dimensions. Fig.12 shows two different process parameters ((a)different combinations of etched process and time, (b)different source/drain metal deposition rate of Au) and their yield distributions at different values.



(a)Etching Time

(b)Source/Drain Metal Deposition Rate of Au

FIGURE 12. Yield Distribution of different process parameters.

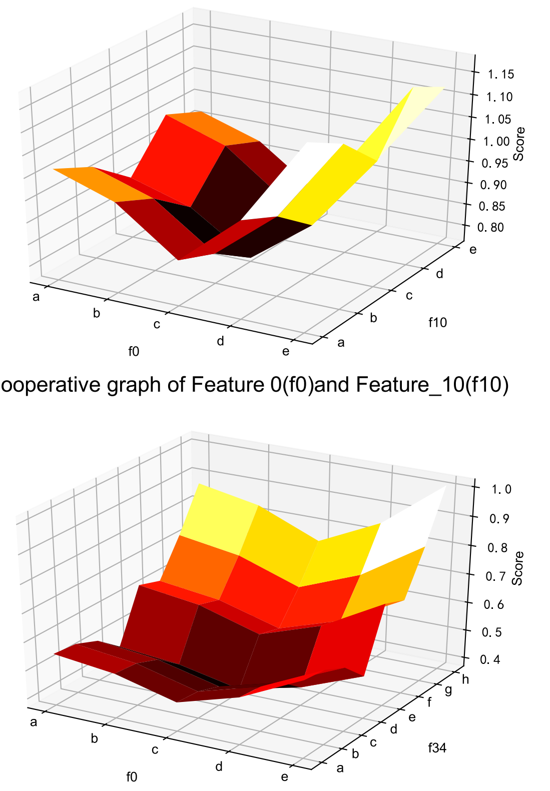
As shown in Fig. 12(a), different etching processes are used in production, and under the same process, the yield prediction score remains unchanged when the total etching duration is less than 180s, and the score remains unchanged when multiple etching processes are used but the total duration does not exceed 180s. Once the total etching time exceeds 180s, the yield prediction score decreases significantly. This is because a long etching time can lead to a residual adhesive denaturation in the channel area that cannot be removed, resulting in a lower yield, so the predicted score is consistent with the actual process.

As shown in Fig.12(b), when source/drain metal deposition rate of Au is 0.5 nm/s, the yield prediction score is the highest, this is because the evaporation of the metal is mostly deposited in the form of clusters of atoms, and the reduction of the rate will improve the density of the clusters, which will lead to the formation of a conductive pathway, which can significantly improve the deposition rate and uniformity, and effectively improve the electrode conductivity in the source/drain region, which in turn improves the yield of the device,. As the deposition rate increases, the score gradually decreases, but when the deposition thickness reaches a certain level, the increase in the thickness of the deposition layer can also improve the conductivity problems that may be caused by uniformity to a certain extent, so there is a certain increase in the data in the last column.

In summary, the atomic-level thickness of 2D semiconductors (e.g., MoS_2) makes them extremely sensitive to etching, and a time deviation of more than $\pm 5\%$ can lead to material tearing or residue, significantly degrading carrier mobility (μ) and threshold voltage (Vth). The model in this paper can guide engineers to avoid over-etching risk zones by accurately predicting the nonlinear effect of etching time on yield. In addition, the metal-semiconductor interface of emerging semiconductors is prone to Fermi energy level pinning, and a fast deposition rate will exacerbate interface defects, leading to a 50% or more drop in I_{on}/I_{off}. The model in this paper optimizes the recommended rate by Bayesian optimization, which can effectively reduce the contact resistance reduction.

Moreover, considering that the actual device manufacturing is done through many processes, we need to observe the joint effect of several process parameters on the yield. Hence, we grouped the most important parameters into a cooperative graph. Fig.13 illustrates the 2-dimensional distribution of these parameters. Subgraphs (a) and (b) depict the yield distribution of two key process parameters in a 3-dimensional plot. The horizontal plane represents two distinct process parameters with varying values, while the vertical axis corresponds to their respective yield scores. Darker regions indicate lower yield scores for a given parameter combination, whereas lighter regions signify higher yields.

We have demonstrated the joint effect of only two parameters on yield; however, in practice, yield is influenced by multiple parameters simultaneously, often numbering in the dozens. Visualizing this multidimensional impact is challenging, so we employ a specialized optimizer in the Yield Tuning experiment to identify the optimal solution across various parameter combinations.



(a) Cooperative graph of Feature 0(f0)and Feature_10(f10)

(b) Cooperative graph of Feature 0(f0)and Feature_34(f34)

FIGURE 13. Cooperative graph in 3D view. The horizontal direction is for two features with different values, and the vertical direction is for scoring.

IV. YIELD TUNING EXPERIMENT AND IMPROVEMENT RESULTS

Once the Yield Diagnosis is done, the Yield Tuning experiment can be performed based on the yield distribution. We gave a new set of process combinations by a Bayesian optimizer and feed it into the Yield Diagnosis mode, repeating the iteration tens of times. During the parameter optimisation process, Bayesian optimisation can autonomously explore high-potential parameter combinations, and the entire iterative process does not require expert intervention, thus further improving the automation level. After running this algorithm, it did find a lot of parameter combinations which were higher than the highest score in the existing data set, the highest of which could reach 140 points. For the case of more than 100 points, it can be explained from 2 aspects. One is that the source of the score is human. The scoring itself is a relative value, and there are no upper or lower bounds. The other one is that it is possible for machine learning models to exceed the defined domain.

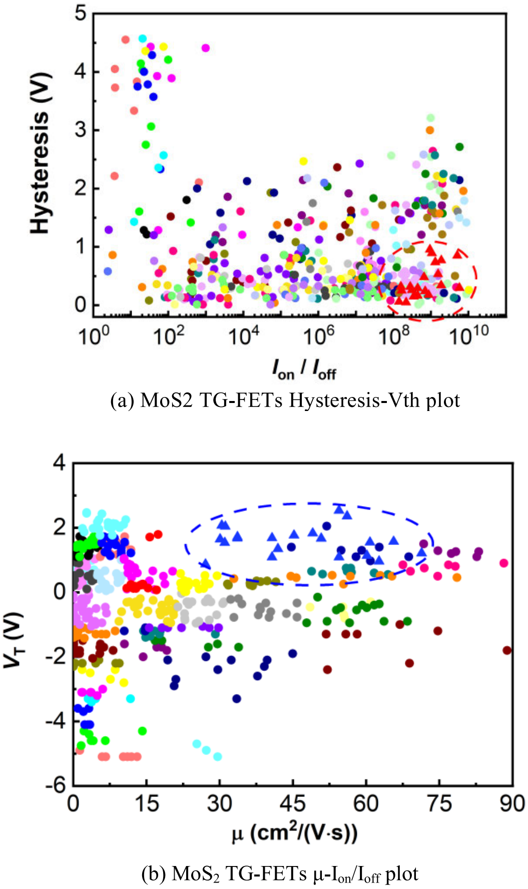
This is called the domain extension in the mathematical field. Whether the extension domain has practical significance depends on the actual situation.

The highest score in the existing data is 86 points, which means we improved the yield by 62%. We will use the prediction results from yield tuning to perform new device experiments and obtain new measurements to verify the correctness of yield tuning. For device researchers, we did not inform them of our predicted scores, while device researchers only knew the suggested parameters for new experiments, so they can evaluate the new experiment according to their criteria and we avoid the influence of psychological factors. The utilization of artificial intelligence as an auxiliary engineering tool within this process serves to complement expert experience, thereby obviating the necessity for human intervention in decision-making. The scalability of the framework in question offers significant advantages over conventional trial-and-error approaches that are predicated on expert experience. The performance of the new device batch after process optimization is demonstrated below.

Due to limited resources, the researchers only counted about 500 sets of devices, including those based on our optimized process combinations, and measured the 4 electrical indicators, including hysteresis, u, I_{on}/I_{off}, and Vth. To show the improved effect of these combinations, two scatter plots were drawn based on the high-performance metrics and the low-power metrics. We set Vth, hysteresis and mobility, I_{on}/I_{off} ratio as vertical and horizontal coordinates for the high-performance and low-power metrics, respectively, label the measurement results of all devices with scatter points on them. Each color represents the same process combination and the devices obtained based on our yield tuning optimization are represented by special triangles.

As can be seen from Fig.14. (a), the optimized devices are basically located in the bottom right corner. The optimized devices achieve an average I_{on}/I_{off} ratio about 2.87 × 109 and a lower hysteresis about 0.41 V which surpass 95.12% and 87.80% of the devices fabricated by other process combinations, respectively, which indicates the less power consumption. Similarly, we adopt μ and Vth as horizontal and vertical coordinates to observe the high-performance of the devices and the results are shown in Fig.14. (b) We can also find the optimized devices marked by special triangles, are basically clustered in the upper right of Fig.14. (b). The figure explicitly shows that they have a higher average Vth (1.59 V) with a higher average μ (46.23 cm2/(V·s)) than 97.56% and 80.50% of the rest device sets, realizing the higher performance compared with other devices. Obviously, these two graphs directly demonstrate the dramatic improvement in device performance that our yield tuning has brought about.

This article mainly analyzes the yield diagnosis and yield tuning of the research phase. There are many similarities between the research phase and the ramp up phase of the factory. Our algorithm is also applicable to semiconductor factories. Basically, both of them have small data volumes. The factory may have more measurement steps and measurement results, which provide more evaluation indicators for machine learning. Engineers can flexibly choose the indicators they want to optimize, either single or ensemble.



(a) MoS2 TG-FETs Hysteresis-Vth plot

(b) MoS_2 TG-FETs μ-I_{on}/I_{off} plot

FIGURE 14. Results of process-optimized devices (indicated by triangles) versus other devices.

If there are some semiconductor devices which are very complicated to manufacture and the process steps are much more than the data used in our experiments, we will not achieve good results by directly applying our algorithm. PCA or other dimensionality reduction algorithm should be used before applying our algorithm, reducing efficient process steps. Furthermore, as the scale of the model increases, the significance of effective data preprocessing and feature engineering becomes paramount. The presence of a high correlation between parameters, as well as issues such as uneven data distribution, necessitates the implementation of rigorous cleaning and dimensionality reduction procedures prior to modelling. These processes ensure the stability and generalisation capability of the model. In conclusion, it is evident that scaling up to larger and more intricate process scenarios demands not only the optimization of algorithms and computational strategies, but also comprehensive adjustments in data management, feature extraction, and modular design to ensure the maintenance of prediction accuracy and system efficiency.

V. CONCLUSION

We use the decision tree-based ensemble learning methods such as XGBoost and Random Forest for Yield Diagnosis for the small amount of data in the device research phase as well as yield ramp-up phase. We also use the Bayesian optimization method for Yield Tuning. Test data is from real semiconductor manufacturing process. Experimental results show that Mean Absolute Error is of no more than 8 % and explained variance is of no less than 0.62 in Yield Prediction test, indicating that the model fits well. Yield Tuning test achieved a maximum yield improvement of 62%. We also fed our results back to the researchers, who combined their experience with the algorithm results to tune the process, remanufactured a batch of devices and measured them, and found an overall improvement in yield. For high performance and low power, the average mobility, Vth and hysteresis, I_{on}/I_{off} ratio of the optimized devices batches surpass 80.50%, 97.56% and 87.80% and 95.12% of the device sets fabricated by other processes combination. Our work has provided an effective platform for the development of new process combinations owing to the obvious improvement of the yield tuning.

ACKNOWLEDGMENT

(Chunshan Wang, Zizhao Ma, Yuxuan Zhu, and Chensheng Jin contributed equally to this work.)

REFERENCES

[1] P. Feng, S.-C. Song, G. Nallapati, J. Zhu, J. Bao, V. Moroz, M. Choi, X.-W. Lin, Q. Lu, B. Colombeau, N. Breil, M. Chudzik, and C. Chidambaram, "Comparative analysis of semiconductor device architectures for 5-nm node and beyond," *IEEE Electron Device Lett.*, vol. 38, no. 12, pp. 1657–1660, Dec. 2017, doi: 10.1109/LED.2017.2769058.

[2] Z. He, "Analysis on the development of semiconductor manufacturing process," *J. Phys., Conf.*, vol. 2295, no. 1, Jun. 2022, Art. no. 012009, doi: 10.1088/1742-6596/2295/1/012009.

[3] W. Cao, J. Kang, D. Sarkar, W. Liu, and K. Banerjee, "2D semiconductor FETs—Projections and design for sub-10 nm VLSI," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3459–3469, Nov. 2015, doi: 10.1109/TED.2015.2443039.

[4] G. Yeap, S. S. Lin, Y. M. Chen, H. L. Shang, P. W. Wang, H. C. Lin, Y. C. Peng, J. Y. Sheu, M. Wang, X. Chen, and B. R. Yang, "5 nm CMOS production technology platform featuring full-fledged EUV, and high mobility channel FinFETs with densest 0.021 μm^2 SRAM cells for mobile SoC and high performance computing applications," presented at the IEDM Tech. Dig., 2019.

[5] Y. Ding, X. Luo, E. Shang, S. Hu, S. Chen, and Y. Zhao, "A device design for 5 nm logic FinFET technology," presented at the China Semicond. Technol. Int. Conf. (CSTIC), 2020.

[6] Y. P. Tsai, Y. H. Chang, J. Wang, D. Trivkovic, K. Ronse, and R. H. Kim, "A yield prediction model and cost of ownership for productivity enhancement beyond imec 5 nm technology node," presented at the DTCO Comput. Patterning, 2022.

[7] H.-W. Xu, Q.-H. Zhang, Y.-N. Sun, Q.-L. Chen, W. Qin, Y.-L. Lv, and J. Zhang, "A fast ramp-up framework for wafer yield improvement in semiconductor manufacturing systems," *J. Manuf. Syst.*, vol. 76, pp. 222–233, Oct. 2024, doi: 10.1016/j.jmsy.2024.07.001.

[8] H.-W. Xu, W. Qin, Y.-L. Lv, and J. Zhang, "Data-driven adaptive virtual metrology for yield prediction in multibatch wafers," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9008–9016, Dec. 2022, doi: 10.1109/TII.2022.3162268.

[9] I. Tirkel, "Yield learning curve models in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 4, pp. 564–571, Nov. 2013, doi: 10.1109/TSM.2013.2272017.

[10] J. Moyne, J. Samantaray, and M. Armacost, "Big data capabilities applied to semiconductor manufacturing advanced process control," *IEEE Trans. Semicond. Manuf.*, vol. 29, no. 4, pp. 283–291, Nov. 2016, doi: 10.1109/TSM.2016.2574130.

[11] C.-F. Chien, C.-Y. Hsu, and P.-N. Chen, ''Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence,'' *Flexible Services Manuf. J.*, vol. 25, no. 3, pp. 367–388, Sep. 2013, doi: 10.1007/s10696-012-9161-4.

[12] B. Lenz and B. Barak, ''Data mining and support vector regression machine learning in semiconductor manufacturing to improve virtual metrology,'' presented at the 46th Hawaii Int. Conf. Syst. Sci., 2013.

[13] T. Chen, ''An ANN approach for modeling the multisource yield learning process with semiconductor manufacturing as an example,'' *Comput. Ind. Eng.*, vol. 103, pp. 98–104, Jan. 2017, doi: 10.1016/j.cie.2016.11.021.

[14] P. Stich, M. Wahl, P. Czerner, C. Weber, and M. Fathi, ''Yield prediction in semiconductor manufacturing using an AI-based cascading classification system,'' in *Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT)*, Jul. 2020, pp. 609–614.

[15] K. B. Irani, J. Cheng, U. M. Fayyad, and Z. Qian, ''Applying machine learning to semiconductor manufacturing,'' *IEEE Exp.*, vol. 8, no. 1, pp. 41–47, Feb. 1993.

[16] F. Bergeret and C. Le Gall, ''Yield improvement using statistical analysis of process dates,'' *IEEE Trans. Semicond. Manuf.*, vol. 16, no. 3, pp. 535–542, Aug. 2003, doi: 10.1109/TSM.2003.815204.

[17] C. Hora, R. Segers, S. Eichenberger, and M. Lousberg, ''An effective diagnosis method to support yield improvement,'' presented at the Int. Test Conf., 2002.

[18] W. Yamwong and T. Achalakul, ''Yield improvement analysis with parameter-screening factorials,'' *Appl. Soft Comput.*, vol. 12, no. 3, pp. 1021–1040, Mar. 2012, doi: 10.1016/j.asoc.2011.11.021.

[19] S. Mao, W. Zhang, Y. Yao, X. Yu, H. Tao, F. Guo, C. Ren, T. Chen, B. Zhang, R. Xu, B. Yan, and Y. Xu, ''A yield-improvement method for millimeter-wave GaN MMIC power amplifier design based on load—Pull analysis,'' *IEEE Trans. Microw. Theory Techn.*, vol. 69, no. 8, pp. 3883–3895, Aug. 2021, doi: 10.1109/TMTT.2021.3088499.

[20] M. Melhem, B. Ananou, M. Ouladsine, and J. Pinaton, ''Regularized regression models to predict the product quality in multistep manufacturing,'' in *Proc. 5th Int. Conf. Syst. Control (ICSC)*, May 2016, pp. 31–36.

[21] Y. Chen, B. Wang, J. Wu, Y. Wu, and C. Chien, ''Big data analytic for multivariate fault detection and classification in semiconductor manufacturing,'' in *Proc. 13th IEEE Conf. Automation Sci. Eng. (CASE)*, Aug. 2017, pp. 731–736, doi: 10.1109/COASE.2017.8256190.

[22] G. Wang, R. M. Hasani, Y. Zhu, and R. Grosu, ''A novel Bayesian network-based fault prognostic method for semiconductor manufacturing process,'' in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2017, pp. 1450–1454.

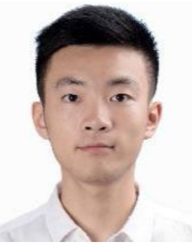
[23] K. B. Lee, S. Cheon, and C. O. Kim, ''A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes,'' *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 2, pp. 135–142, May 2017.

[24] Z. Ahmed, A. Afzalian, T. Schram, D. Jang, D. Verreck, Q. Smets, P. Schuddinck, B. Chehab, S. Sutar, G. Arutchelvan, A. Soussou, I. Asselberghs, A. Spessot, I. P. Radu, B. Parvais, J. Ryckaert, and M. H. Na, ''Introducing 2D-FETs in device scaling roadmap using DTCO,'' presented at the IEDM Tech. Dig., 2020.

[25] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, ''Taking the human out of the loop: A review of Bayesian optimization,'' *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016.

[26] J. Snoek, H. Larochelle, and R. P. Adams, ''Practical Bayesian optimization of machine learning algorithms,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.

[27] P. I. Frazier, ''Bayesian optimization,'' in *Recent Advances in Optimization and Modeling of Contemporary Problems*. Informs, 2018, pp. 255–278.



CHUNSHAN WANG received the B.S. degree from the College of Electronic Science and Engineering, Jilin University, China. He is currently pursuing the Ph.D. degree with the School of Microelectronics, Fudan University. His research interests include advanced memory circuit design, domain-specific circuits, and AI accelerator design.

ZIZHAO MA received the bachelor's degree in microelectronics from Xidian University, in 2016. He is currently pursuing the Ph.D. degree with Fudan University, specializing in novel compute-in-memory techniques. His primary research interests include high-precision compute-in-memory, emerging memory circuit, and compute-in-memory-based signal processors.

YUXUAN ZHU received the B.S. degree from Wuhan University of Technology, in 2021. He is currently pursuing the master's degree with the School of Microelectronics, Fudan University, China. His main research interests include electronic devices and advanced processes of 2D material.

CHENSHENG JIN received the B.S. degree from Wuhan University of Science and Technology. He is currently pursuing the M.S. degree in electronic science and technology with the School of Microelectronics, Fudan University. His research interests include new energy system design and compute-in-memory hardware design.

DONGYU CHEN received the B.S. degree in electronic science and technology from Harbin Institute of Technology, Weihai, in 2022. He is currently pursuing the M.S. degree with the State Key Laboratory of Integrated Chips and Systems, Fudan University, China. His research interests include compute-in-memory (CIM), artificial intelligence (AI), and AI accelerators.

CHUXIN ZHANG received the dual B.S. degree in circuit science and engineering, in 2023. She is currently pursuing the M.S. degree with the State Key Laboratory of Integrated Chips and Systems, Fudan University, China. Her research interests include compute-in-memory (CIM), artificial intelligence (AI), and AI accelerator design.

YINING CHEN (Associate Member, IEEE) received the bachelor's degree in electronics and information engineering from the School of Electrical Engineering, Zhejiang University, Zhejiang, China, in 2004, the master's degree in materials for microelectronics from K. U. Leuven, Belgium, in 2005, and the Ph.D. degree from the Division of Microelectronics, School of EEE, Nanyang Technological University, Singapore, in 2011.

In 2021, he joined the College of Integrated Circuits, Zhejiang University, and the distinguished Researcher of the School. His current research interests include integrated circuits process development, integrated circuits yield enhancement, big data and machine learning in semiconductor manufacture, DTCO and IC reliability study, and failure analysis.

WENZHONG BAO received the Ph.D. degree from the Department of Physics and Astronomy, University of California, Riverside, in 2011. He subsequently held postdoctoral positions with the University of Maryland College Park, from 2011 to 2014, and the King Abdullah University of Science and Technology, from 2014 to 2015. He is currently a Full Professor with the School of Microelectronics, Fudan University. His current research interests involve emerging semiconductors and their applications in next-generation electrical, optoelectrical, and energy devices. He was awarded the 2016 International Union of Pure and Applied Physics (IUPAP) Young Scientist Prize (C10) and the 2017 Hong Kong Qiushi Outstanding Young Scientist Prize.

YUFENG XIE (Member, IEEE) received the bachelor's degree in electronics science and technology from Xi'an Jiaotong University, in 2002, and the Ph.D. degree in microelectronics from Tsinghua University, in 2008. Since 2008, she has been with the School of Microelectronics, Fudan University, as a Lecturer, an Associate Professor, and a Professor. Her research interests include advanced memory circuits, computing-in-memory circuits, hardware security of storage, and artificial intelligence hardware and applications.