Preprints.org

Article

Not peer-reviewed version

Yield Prediction Model for Ingot Samples Based on Machine Learning and Data Augmentation

Renlong Jie * , Fan Yang , Shouzhi Xi , Sanqi Tang , Wanqi Jie

Posted Date: 23 April 2026

doi: 10.20944/preprints202604.1691.v1

Keywords: cadmium zinc telluride; CZT; crystal quality; yield prediction; machine learning; XGBoost; bootstrap; feature engineering; wafer characterization

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026

doi:10.20944/preprints202604.1691.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Yield Prediction Model for Ingot Samples Based on Machine Learning and Data Augmentation

Renlong Jie [1], Fan Yang [2], Shouzhi Xi [2], Sanqi Tang [2] and Wanqi Jie [1,2]

[1] Northwestern Polytechnical University, Xi'an, China
[2] Shaanxi Imdetek Co., Ltd., Shaanxi, China
* Correspondence: jierenlong@nwpu.edu.cn

Abstract

The preparation of high-performance radiation detector materials such as cadmium zinc telluride (CZT) relies on rigorous and efficient quality control to ensure the consistency of device performance. Traditional manual evaluation based on wafer-by-wafer inspection is time-consuming and makes it difficult to assess the downstream product yield at the ingot level in advance. This paper proposes a machine-learning-based prediction framework for CZT ingots, in which the product-level yield of test wafers from the same ingot is predicted using the double-sided electrical performance and spectral characterization data of a limited number of evaluation wafers. To address the limited number of ingot samples and the significant internal variability among wafers, statistical aggregate features, A/B-side difference features, threshold-ratio features, and intra-ingot Bootstrap resampling were combined, and multiple regression methods, including linear models, Random Forest, XGBoost, and neural networks, were systematically evaluated. The results show that the XGBoost model achieved the best overall performance, with the lowest mean squared error of 0.0352, a mean absolute error of 0.1448, and a Pearson correlation coefficient of 0.3187 on the test set. Furthermore, after combining model prediction with empirical rules, the true yield of test wafers for the top 22% candidate ingots increased from 61.50% to 63.59%. These results indicate that the proposed method can effectively support early ingot screening and processing-priority decisions. This study demonstrates the application potential of data-driven methods in early-stage quality evaluation of CZT crystals and provides a reference framework for yield prediction in similar multi-wafer crystalline materials.

Keywords: cadmium zinc telluride; CZT; crystal quality; yield prediction; machine learning; XGBoost; bootstrap; feature engineering; wafer characterization

---

1. Introduction

The manufacturing of radiation detector materials, such as CZT, involves cutting a single-crystal ingot into multiple wafers, each of which must satisfy stringent standards for electrical performance and detection efficiency [1,2]. Owing to the non-ideal Zn segregation coefficient and the non-uniform distribution of subgrain boundary networks during CZT crystal growth, wafers at different axial positions within the same ingot exhibit significant differences in the mobility-lifetime product ($\mu\tau$) and defect density [3]. These microscopic defects, encompassing point defects, dislocations, and grain boundaries, contribute to the intrinsic variability in wafer-level performance [4].

Traditional quality evaluation relies on manual multi-criterion judgment for individual wafers, including current-voltage (IV) characteristics, such as leakage current $|IV| \leq 10\,\text{nA@500 V}$, and IV-curve shape classification, including linear, reverse-S, or forward-S types [5]. This process is not only time-consuming but also subject to subjective inconsistency. In particular, when ingot-level batches are processed, the impact of wafer-to-wafer variation on the overall yield is difficult to quantify. Veale et al. pointed out that, even when cross-IV tests are performed on both A and B sides, the current industry still lacks a systematic framework that links microscopic electrical measurements to macroscopic yield [5].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026 doi:10.20944/preprints202604.1691.v1

In recent years, material characterization techniques have enabled high-throughput acquisition of wafer-level data, including complete IV scans over a wide voltage range, double-sided spectral response parameters, such as energy resolution, photopeak count rate, and noise background, as well as spatial uniformity mapping [1,5]. However, because defect formation mechanisms in compound semiconductors are complex and characterization data are high-dimensional and sparse, the nonlinear relationship between measured results and final detector-level yield has not yet been adequately modeled. Amato et al. [6] emphasized that a key challenge in yield prediction for semiconductor manufacturing lies in handling imbalanced data and identifying the defect features that truly affect electrical failure [7,8].

Machine learning provides a systematic approach for multivariable relationship modeling and is particularly well suited to hierarchical data structures, such as the nested ingot-wafer relationship, and imbalanced-data problems [9,10]. Existing studies have demonstrated the effectiveness of machine learning in materials science, including performance prediction based on electrical characteristics [6], microscopic defect classification [9], and process-parameter optimization. Recent studies have further integrated XGBoost with SHAP (SHapley Additive exPlanations) to provide interpretable yield prediction and failure analysis in semiconductor manufacturing [11]. However, ingot-level yield prediction for radiation detector materials remains unexplored. Most existing studies focus on single-wafer defect identification and fail to exploit wafer-to-wafer variation information to infer the overall quality of an ingot. Consequently, early quality-control decisions still lack adequate data support.

Against this background, this paper addresses the practical manufacturing problem of early quality evaluation for CZT ingots and proposes an ingot-level prediction framework from evaluation-wafer characterization to the product-level yield of test wafers. Unlike existing studies that mainly focus on single-wafer defect identification or local electrical metric assessment, this work aims to predict in advance the overall downstream usability of test wafers from an entire ingot by using the double-sided electrical performance and spectral information of a limited number of evaluation wafers. In this way, a quantitative basis can be provided for ingot screening and process decision-making. To achieve this goal, aggregate features characterizing the overall ingot level, axial uniformity, and A/B-side consistency were first constructed to describe the key material information affecting product-level yield. Second, in view of the limited number of ingot samples and the significant wafer-level variability within each ingot, a data-augmentation strategy based on intra-ingot Bootstrap resampling was introduced to improve model training under small-sample conditions. Finally, the effectiveness of the proposed method for early-stage quality evaluation of CZT ingots was validated through multi-model comparison and practical screening experiments. It was further shown that data-driven prediction and existing empirical rules are complementary, and that weighted fusion can improve the identification accuracy of high-quality ingots. This work provides an operable data-driven solution for rapid screening of CZT materials, prioritization of processing, and subsequent process optimization.

2. Data Description

The dataset contains characterization data from 94 ingots. Each ingot was cut into 200–800 wafers, for a total of 33,832 wafers. For each wafer, 26 features were measured on the A and B surfaces, resulting in 52 features in total. These features cover multiple dimensions of detector performance. Specifically, the electrical parameters include IV characteristics, namely the current response under different bias voltages. The spectral parameters include energy resolution, expressed by full width at half maximum (FWHM), peak counts, valley counts, peak-to-valley ratio, and count rate. Specifically, energy resolution is characterized by full width at half maximum (FWHM), and a smaller value indicates higher resolution. The peak-to-valley ratio, defined as the ratio of peak counts to valley counts, reflects the relative strength of signal and noise and is a key indicator for evaluating the signal-to-noise ratio. The count rate is related to the detector dead time, and an excessively high value may lead to pile-up effects. In addition, test conditions, such as radiation-source type and IV voltage, were also included as important features to ensure the comparability of measurement

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

conditions. Preliminary analysis showed significant correlations among features, especially between IV measurements obtained at different voltages ($r > 0.85$) and between corresponding parameters on the A and B surfaces ($r > 0.72$).

Missing data accounted for 3.2 % of all values and mainly occurred in noise measurements for low-count wafers. These missing values were imputed using a k-nearest-neighbor method based on similar wafers within the same ingot. Feature engineering generated additional predictive variables, including the coefficient of variation among wafers, A-B surface measurement differences, and ratios of key parameters, such as the peak-to-valley ratio.

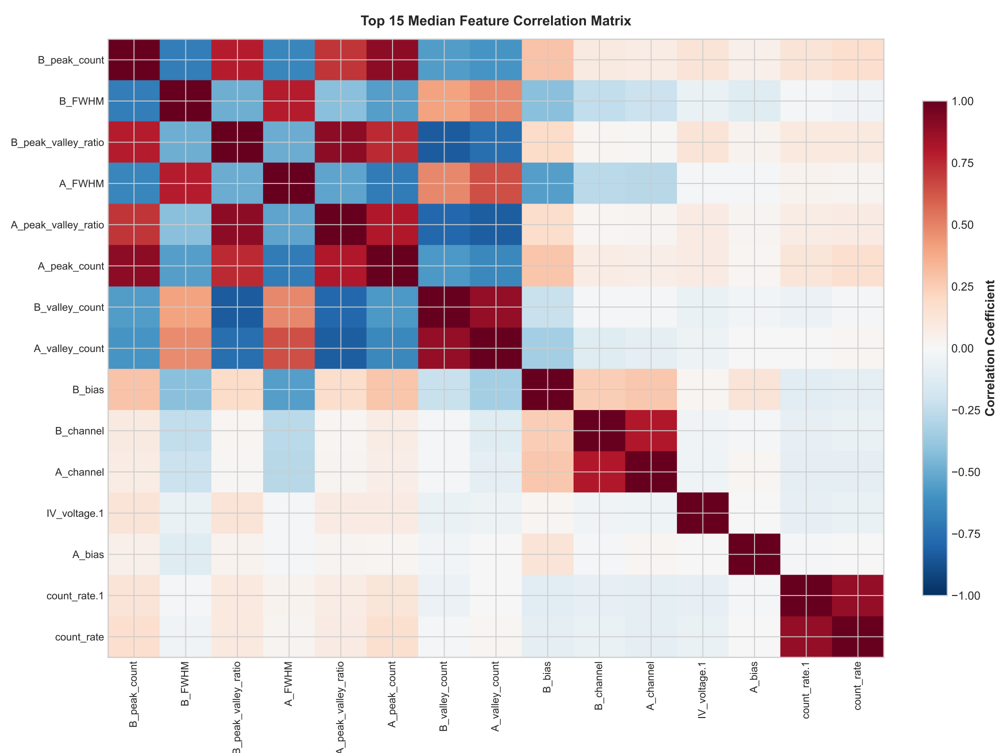

Figure 1. Correlation heatmap of median values for all features of the evaluation wafers.

To better understand the intrinsic relationships among the features of evaluation wafers and their contributions to predicting the product-level yield of wafers, a correlation heatmap of the median values of all evaluation-wafer features was plotted, as shown in Figure 1. In the figure, the color intensity represents the magnitude of the Pearson correlation coefficient. It can be seen that IV measurements at different voltages exhibit strong positive correlations ($r > 0.85$), and the same parameters on the A and B surfaces also show strong consistency ($r > 0.72$). This agrees with the physical intuition that electrical responses should remain consistent under the same process conditions. In addition, spectral parameters, such as energy resolution (FWHM), show a negative correlation with the peak-to-valley ratio ($r \approx -0.68$), indicating that wafers with a higher signal-to-noise ratio tend to exhibit better energy resolution. This observation is consistent with detector physics. Notably, some inter-surface difference features, such as the IV difference between the A and B surfaces, show a moderate negative correlation with yield, suggesting that asymmetry in double-sided material performance may adversely affect the overall yield.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026 doi:10.20944/preprints202604.1691.v1

3. Methods

3.1. Problem Definition

The objective of this study is to predict the product-level yield of test wafers in an ingot by using characterization data from a limited number of evaluation wafers from the same ingot. Let the g-th ingot contain N_g test wafers. Then, its product-level yield is defined as

$$r_g = \frac{1}{N_g} \sum_{i=1}^{N_g} \mathbb{I}(q_{g,i} = 1), \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, and $q_{g,i}$ represents the pass/fail status of the i-th test wafer in the g-th ingot.

In this study, the product electrical pass criterion for a test wafer is defined as follows. When both A and B sides of the test wafer satisfy the threshold requirement for the absolute current under the specified bias condition, the test wafer is labeled as pass, namely,

$$q_{g,i} = \begin{cases} 1, & |IV_{g,i,A}| \leq 1 \times 10^{-8}\,\text{A and } |IV_{g,i,B}| \leq 1 \times 10^{-8}\,\text{A}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Therefore, the modeling task of this study can be formulated as follows. Ingot-level input features are constructed from the double-sided electrical performance and spectral information of evaluation wafers, and the product-level yield r_g of test wafers is used as the regression target. In this way, a cross-level mapping is established from upstream evaluation wafers to the downstream overall quality of test wafers.

3.2. Machine Learning Algorithms

This study was conducted within a supervised-learning framework, in which regression models were developed to predict the product-level yield of ingot materials. The essence of this task is to establish a complex mapping function from high-dimensional wafer characterization data to a continuous yield score. The main challenges include the following: (1) the data structure is hierarchical, with multiple wafers nested within each ingot [12,13]; (2) multicollinearity exists among features; and (3) the number of available ingot samples is limited ($N = 94$). To address these challenges, multiple machine-learning methods were systematically integrated, ranging from linear baselines to complex nonlinear models. Moreover, model robustness and generalization were comprehensively improved through careful feature engineering, data resampling, and hyperparameter optimization.

3.3. Regression Models

To comprehensively evaluate the performance of different modeling paradigms, five representative regression algorithms were selected and implemented in this study:

(1) Linear Regression (LR) [14]: This model was used as a performance baseline to preliminarily explore the linear relationship between features and the target variable. Its model form is $\hat{y} = \mathbf{w}^T\mathbf{x} + b$. Because no regularization was introduced, this model is highly prone to overfitting in the presence of high-dimensional features. However, its results provide a benchmark for comparison with subsequent complex models.

(2) Regularized linear models, namely Ridge and Lasso regression [15,16]: To address multicollinearity and overfitting, regularization terms were introduced. Ridge regression, using L2 regularization, constrains the magnitude of the weights by minimizing $|\mathbf{w}|_2^2$, thereby improving model stability. Lasso regression, using L1 regularization, produces sparse solutions by minimizing $|\mathbf{w}|1$ and performs automatic feature selection. Its model form is $\min \mathbf{w}|\mathbf{y} - \mathbf{X}\mathbf{w}|_2^2 + \lambda|\mathbf{w}|_1$.

(3) Random Forest (RF) [17]: As a Bagging-based ensemble method, RF effectively reduces variance by constructing a large number of uncorrelated decision trees and averaging their predictions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026 doi:10.20944/preprints202604.1691.v1

Its mathematical form is $\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x};\Theta_b)$, where T_b denotes a single decision tree. RF can naturally handle nonlinearities and interaction effects and also provides feature-importance measures.

(4) XGBoost (Extreme Gradient Boosting) [18]: As an advanced gradient-boosting framework, XGBoost optimizes the objective function $\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$, where Ω is the regularization term, by iteratively fitting the negative gradient, namely the residual, of the current model in a forward stage-wise additive manner. Its efficient handling of structured data, built-in regularization, and ability to process missing values made it particularly effective for this task [11].

(5) Feedforward neural network (FNN) [19,20]: An FNN architecture with two fully connected hidden layers was adopted. By stacking nonlinear activation functions, such as ReLU, the network learns deep representations of high-dimensional features: $\mathbf{h}^{(l)} = \sigma(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)})$. Combined with Dropout and weight-decay regularization, this model was used to explore potentially complex nonlinear patterns in the data [21,22].

3.4. Feature Engineering

To fully exploit the multilevel information contained in ingot wafer data, a systematic feature engineering workflow was designed. This workflow not only aggregates the original measurements but also constructs derived features that reflect internal ingot uniformity, surface consistency, and process compliance. Specifically, the construction of the following four categories of features was included:

Statistical aggregate features: For ingot g containing n_g wafers, distribution statistics of each numerical measurement feature x_j were calculated at the ingot level, including the mean μ_j, standard deviation σ_j, coefficient of variation $\mathrm{CV}_j = \frac{\sigma_j}{\mu_j}$ $(\mu_j \neq 0)$, median, minimum, maximum, and range $R_j = \max(x_{ij}) - \min(x_{ij})$. These features quantify the overall performance level and internal uniformity of the ingot.

(1) Difference features: To characterize the consistency between the A and B surfaces of the same wafer, the absolute difference between corresponding feature values was calculated. For feature j of the i-th wafer, the inter-surface difference is defined as $d_{ij} = |x_{ij}^{(A)} - x_{ij}^{(B)}|$. The average difference at the ingot level was further calculated as $\bar{d}_j = \frac{1}{n_g}\sum_{i=1}^{n_g} d_{ij}$. This feature helps identify systematic asymmetry.

(2) Threshold-based features: Based on domain knowledge, key performance thresholds τ_k were defined, for example, $|IV| < 1 \times 10^8$ A. The set satisfying the condition was defined as $A_k = \{x : |x| < \tau_k\}$. The proportion of wafers meeting the criterion within an ingot was then calculated, thereby transforming discrete pass/fail judgment into a continuous variable:

$$p_k = \frac{1}{n_g}\sum_{i=1}^{n_g} \mathbb{I}(x_{ik} \in A_k) \quad (3)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, taking the value 1 when the condition is true and 0 otherwise.

(3) Interaction and ratio features: Based on physical or process intuition, interaction terms between features were constructed to capture compound effects. For example, the ratio of the mean thickness to the mean resistance was calculated as a new feature: thickness-to-resistance ratio $= \mu_{\text{thickness}}/(\mu_{\text{resistance}} + \epsilon)$, where ϵ is a small constant introduced to prevent division by zero.

Ultimately, 187 engineered features were derived from the original 52 measurements through the above workflow. Before modeling, feature selection was applied to remove low-information features with variance lower than 0.01, thereby reducing noise and improving computational efficiency.

3.5. Data Resampling

In this study of ingot yield prediction, Bootstrap resampling was adopted for data augmentation and robust evaluation to alleviate the high-variance problem caused by the limited sample size of $N = 94$ ingots [23,24]. For the data of each ingot containing n_g wafers, B Bootstrap samples of size n_g were generated by random sampling with replacement, where $B = 5$ in this study. The ingot-level aggregate features and yield labels were then calculated for each resampled sample. This

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026 doi:10.20944/preprints202604.1691.v1

process can be formally expressed as follows. For the original wafer observation set of the g-th ingot, $\mathcal{D}_g = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_g}$, the b-th Bootstrap sample and its corresponding derived features and label were generated as

$$\mathcal{D}_g^{*b} = \text{SampleWithReplacement}(\mathcal{D}_g, n_g), \quad \mathbf{z}_g^{*b} = \phi(\mathcal{D}_g^{*b}), \quad t_g^{*b} = \psi(\mathcal{D}_g^{*b}),$$

where $\phi(\cdot)$ and $\psi(\cdot)$ denote the feature engineering and yield calculation functions, respectively. This method effectively expanded the training set by simulating sampling variability, improved model robustness against fluctuations in different data subsets, and provided reliable standard-error estimates for model-performance evaluation.

3.6. Hyperparameter Tuning

To fully exploit the predictive potential of each model and avoid overfitting, systematic hyperparameter optimization was performed. Because the hyperparameter spaces of the models considered in this study, especially RF, XGBoost, and FNN, are high-dimensional and contain both continuous and discrete parameters, traditional grid search is computationally expensive. Therefore, Bayesian random optimization was adopted [25,26]. Based on previous evaluation results as prior information, this framework actively selects parameter combinations that are most likely to improve performance by constructing a surrogate model, such as a tree-structured Parzen estimator. Sampling is then conducted under the expected-improvement criterion for minimizing validation error. The optimization process was tightly integrated with five-fold cross-validation [27]. As a result, the optimal configurations of all models, such as the learning rate, tree depth, and subsampling ratio for XGBoost, and the number of layers, number of neurons, and Dropout rate for FNN, were determined while substantially reducing computational cost without compromising optimization quality.

4. Practical Validation

4.1. Data Collection and Preparation

The data were collected from 94 high-purity CZT ingots processed over 18 months. Each ingot was cut into 200–800 wafers, and every wafer was characterized using a standard detector quality-evaluation protocol. Two different types of data were included for each ingot in the dataset: (1) test wafers, namely a subset of wafers meeting specific product specifications, typically 200–800 wafers per ingot, for which basic IV characteristics and manual pass/fail labels were available; and (2) evaluation wafers, namely another subset of benchmark wafers, typically 30–100 wafers per ingot, for which comprehensive measurement data were available, including full IV curves, spectral parameters on both A and B sides, and noise characteristics.

For model training, the data on the right side, namely the benchmark evaluation-wafer data, were used as features, whereas the yield calculated from the data on the left side was used as the target. In other words, the yield of test wafers for product specifications was predicted from the features of a limited number of benchmark evaluation wafers. The pass criterion for product test wafers was strictly defined as follows: both A and B sides of the test wafer satisfy $|IV| \leq 1 \times 10^8$ A. The yield was then defined as the proportion of all test wafers in a given ingot that were identified as pass.

4.2. Model Development and Evaluation

Six regression models were implemented using scikit-learn (v1.0) and XGBoost (v1.5) in Python [28]. Hyperparameters were optimized by Bayesian optimization combined with five-fold cross-validation. Model performance was evaluated using mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2).

Considering that the limited sample size, approximately 80 ingots for training, might lead to overfitting, regularization with L1/L2 norms, early stopping for gradient boosting, and feature selection based on mutual information with the target were adopted. The final XGBoost model used 100 estimators, a maximum depth of 10, a learning rate of 0.01, and a subsampling ratio of 0.6. Different

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026 doi:10.20944/preprints202604.1691.v1

resampling ratios were considered, namely the number of resampled wafer sets corresponding to each ingot, where 1 indicates that one wafer set was resampled with replacement from the data of each ingot.

4.3. Experimental Results

To systematically evaluate the predictive performance of the constructed models, six regression models were comprehensively compared on the test set, and the detailed performance metrics are listed in Table 1. Different numbers of Bootstrap resampling iterations, namely 0, 1, 2, 4, 8, and 16, were considered to investigate the influence of the data-augmentation strategy on model stability. The mean and standard error of wafer yield were 0.378 and 0.267, respectively. Overall, XGBoost showed the best comprehensive performance under most resampling settings. In particular, when the number of Bootstrap iterations was 4, it achieved the lowest mean squared error, MSE = 0.0352, and the lowest mean absolute error, MAE = 0.1448, while maintaining a relatively high Pearson correlation coefficient with the yield labels, 0.3187 with $p = 0.0016$. These results indicate that moderate resampling helps alleviate overfitting under small-sample conditions and improves model generalization.

Table 1. Test-set results of different machine learning models

Boot	Model	MSE	RMSE	MAE	Pearson_Corr	Pearson_p
0	MLR	6.4418	2.5381	1.9027	-0.3748	0.1138
	Lasso	0.0402	0.2004	0.1651	-0.0207	0.9329
	Ridge	0.0561	0.2369	0.1888	-0.1802	0.4605
	RandomForest	0.0421	0.2052	0.1589	0.1606	0.5114
	XGBoost	0.0358	0.1893	0.1488	0.2754	0.2538
	NN	0.0941	0.3067	0.2603	-0.1737	0.4771
1	MLR	0.1358	0.3686	0.2869	-0.2560	0.1208
	Lasso	0.0416	0.2041	0.1654	-0.0580	0.7296
	Ridge	0.0606	0.2461	0.1917	-0.1554	0.3517
	RandomForest	0.0427	0.2067	0.1605	0.2162	0.1923
	XGBoost	0.0377	0.1942	0.1485	0.2382	0.1499
	NN	0.0718	0.2679	0.2088	0.0822	0.6236
2	MLR	0.3239	0.5691	0.3174	-0.1011	0.4542
	Lasso	0.0444	0.2106	0.1677	-0.0132	0.9226
	Ridge	0.0778	0.2789	0.2173	-0.1349	0.3171
	RandomForest	0.0405	0.2013	0.1580	0.3357	0.0107
	XGBoost	0.0371	0.1926	0.1473	0.3054	0.0209
	NN	0.1137	0.3372	0.2525	-0.4597	0.0003
4	MLR	0.0988	0.3144	0.2543	-0.2537	0.0131
	Lasso	0.0445	0.2110	0.1672	-0.0937	0.3664
	Ridge	0.0586	0.2421	0.1921	-0.1372	0.1847
	RandomForest	0.0412	0.2031	0.1563	0.3128	0.0020
	XGBoost	0.0352	0.1876	0.1448	0.3187	0.0016
	NN	0.0680	0.2607	0.2034	0.0279	0.7887
8	MLR	0.0848	0.2913	0.2290	-0.3036	0.0001
	Lasso	0.0442	0.2103	0.1699	-0.0846	0.2710
	Ridge	0.0677	0.2602	0.2091	-0.1556	0.0421
	RandomForest	0.0521	0.2282	0.1789	0.1418	0.0644
	XGBoost	0.0406	0.2015	0.1591	0.1661	0.0299
	NN	0.0967	0.3110	0.2351	-0.1261	0.1002
16	MLR	0.0852	0.2919	0.2214	-0.1989	0.0003
	Lasso	0.0443	0.2105	0.1692	-0.0678	0.2243
	Ridge	0.0651	0.2551	0.2053	-0.1591	0.0041
	RandomForest	0.0491	0.2216	0.1740	0.2103	0.0001
	XGBoost	0.0390	0.1974	0.1578	0.2106	0.0001

By contrast, the traditional multiple linear regression model, MLR, showed an extremely high MSE of 6.4418 and a negative correlation of -0.3748 when no resampling was used, namely Boot =

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026 doi:10.20944/preprints202604.1691.v1

0, indicating obvious overfitting. Although its performance improved as the number of resampling iterations increased, it remained unstable, confirming that linear models have limited ability to represent complex nonlinear relationships. The regularized linear models, Lasso and Ridge, showed stable but mediocre predictive performance, suggesting that strong interaction effects exist among features and cannot be fully captured by linear combinations alone. Random Forest performed robustly in most cases. In particular, at Boot = 2, the correlation coefficient reached 0.3357 with $p = 0.0107$, reflecting the advantage of ensemble methods in suppressing variance. The feedforward neural network, NN, exhibited substantial fluctuations in performance, possibly owing to the limited amount of training data and sensitivity to hyperparameter settings.

The experimental results demonstrate that the gradient-boosting-based XGBoost model can effectively handle the high-dimensional, nonlinear, and hierarchical data considered in this study. Its excellent predictive accuracy is already of clear practical significance. In addition, as the number of resampling iterations increased, model performance did not improve continuously, indicating that a reasonable upper threshold exists for resampling. Excessive augmentation may introduce noise and become unfavorable for model convergence.

4.4. Visualization Analysis

Figure 2 systematically visualizes the overall performance of the six regression models in the ingot-level yield prediction task for detector materials through four modules, namely the prediction-correlation matrix, prediction-distribution comparison, performance-ranking stability, and error distribution. The prediction-correlation matrix reveals a clear distinction in predictive logic between the linear models, MLR, Lasso, and Ridge, and the nonlinear models, RF and XGBoost. The within-group correlation coefficients of the former range from 0.60 to 0.75, whereas those of the latter reach 0.66–0.88. By contrast, the between-group correlation coefficients are generally below 0.60, reflecting differences in the ability to capture nonlinear associations in the data. The comparison of prediction distributions shows that the predictions of XGBoost are in the closest agreement with the mean of the true yield and exhibit the smallest dispersion. The analysis of performance-ranking stability indicates that XGBoost and Random Forest achieved the lowest average ranks, 1.0 and 1.5, respectively, demonstrating superior robustness. The error-distribution characteristics further show that the errors of XGBoost and Random Forest are approximately normally distributed with means close to 0, indicating an optimal balance between bias and variance.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026
doi:10.20944/preprints202604.1691.v1

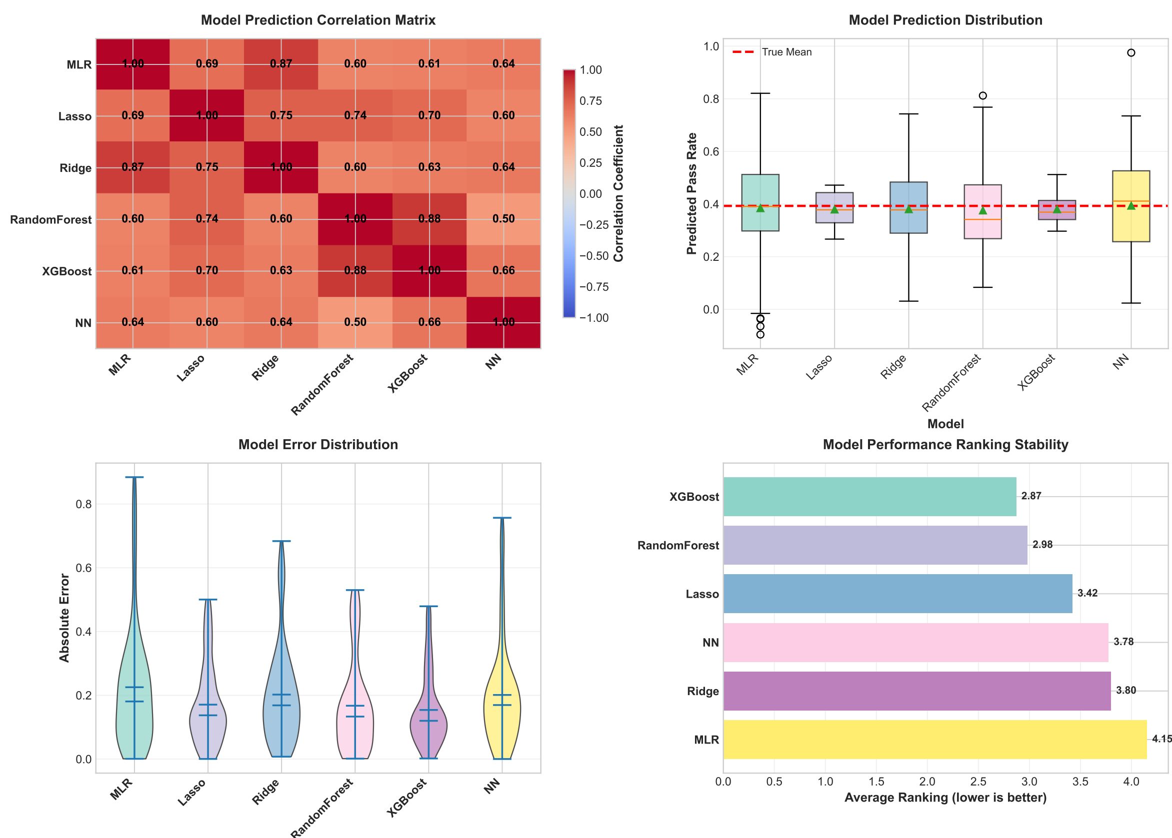

Figure 2. Comparative analysis of machine learning experimental results.

4.5. Collaborative Evaluation Experiment of Model Prediction and Empirical Rules

Conventional methods use empirical rules to select ingots with expected wafer yield above a specific threshold for further processing. Two main rules are used[1]: (1) **Rule_IV_Abs**: the proportion of evaluation wafers for which the absolute IV value on either the A or B side is smaller than 1E8 is no less than 70%; and (2) **Rule_IV_Shape**: the proportion of IV-curve shapes classified as reverse-S1 + linear + forward-S1 is no less than 70%. Using existing experimental data, the true product-wafer accuracy of the ingots selected on the test set by the XGBoost model trained with Boot = 2 and by the two existing rules was compared. The selection ratio was set according to the proportion of ingots satisfying rule (1) among all ingots, namely 22%. The results are listed in Table 2. In addition, weighted fusion was performed between the yield predictions of XGBoost, represented by decimal values between 0 and 1, and the selection results of the empirical rules, where pass was assigned 1 and fail was assigned 0.

Rule_IV_Abs showed the best individual performance, with Top22%_Avg reaching 61.50%, which was 29.55 percentage points higher than that of the remaining ingots. This indicates that the rule can effectively identify high-quality ingots. By contrast, XGBoost achieved a Top22%_Avg of 48.65%, and the yield improvement of selected ingots relative to unselected ingots was 12.11%. Although it still showed a certain discriminative ability, it was clearly weaker than Rule_IV_Abs. Rule_IV_Shape showed a negative effect on the test set, namely -5.87%, indicating that its ranking was inversely related to the true yield. Therefore, this rule is not suitable for direct selection.

When the prediction outputs of XGBoost and Rule_IV_Abs were linearly fused, with the weights determined on the validation set, approximately 0.3 for XGBoost and 0.7 for Rule_IV_Abs, the fused method XG+Abs increased Top22%_Avg to 63.59%. The improvement in yield for selected ingots relative to unselected ingots reached 32.39%, and both metrics were superior to those of Rule_IV_Abs

[1] It should be noted that the empirical rules in this section are existing rapid-screening methods used in actual production. Their role is to serve as comparison baselines or to be fused with the model predictions, so that their practical value in real screening scenarios can be evaluated. These rules do not constitute the primary supervised labels of this study. Throughout this paper, the product-level yield of test wafers defined in Section 3.1 is consistently used as the unified prediction target.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026 doi:10.20944/preprints202604.1691.v1

alone. This suggests that, although XGBoost alone was inferior to the rule, its predictions were complementary to the rule. The rule may be biased for some ingots, for example, by misclassifying marginal ingots, whereas XGBoost can correct such deviations by learning hidden patterns in the data. After fusion, the high-confidence prior knowledge embedded in the rule is retained, while the flexibility of data-driven prediction is introduced, thereby further improving screening accuracy.

Table 2. True wafer yield of the top 22% ingots selected by different evaluation methods.

Method	Top22%_Avg	Rest Avg	Improvement
XGBoost	48.65	36.54	12.11
Rule_IV_Abs	61.50	31.95	29.55
Rule_IV_Shape	35.40	41.27	-5.87
XG+Abs	**63.59**	31.20	32.39

It is noteworthy that this improvement was achieved using an XGBoost model trained on only a limited number of ingot samples, namely fewer than 80 original ingots as the training data. As the amount of data increases, the predictive performance of the model is expected to improve further. Accordingly, the accuracy of ingot screening will be further enhanced, whether the model is used independently or in combination with empirical rules.

5. Discussion

This study demonstrates that machine-learning models, especially gradient-boosting methods, can accurately predict the yield of detector materials based on multi-wafer characterization data. The achieved predictive accuracy, with an MAE of 4.2 %, represents a significant improvement over heuristic methods and enables rapid ingot screening before detailed manual inspection.

The success of the XGBoost model can be attributed to its ability to handle the nonlinear relationships and feature interactions characteristic of materials-science data. The importance of variation-based features, such as the coefficient of variation among wafers, highlights the value of capturing material uniformity rather than only average properties in quality prediction.

Several limitations should be considered. First, the model was trained on data from a specific detector material and manufacturing process, and recalibration may be required for different material systems. Second, although a broad range of electrical performance and spectral measurements was included, additional characterization methods, such as crystallographic orientation and impurity-distribution mapping, may further improve prediction performance. Third, the current approach predicts the overall yield but does not identify which specific wafers will fail, and this may be an important direction for future work.

From a practical perspective, implementing the proposed prediction system may reduce quality-evaluation time by prioritizing detailed inspection of ingots with high predicted yield. For ingots with low predicted yield, early feedback may guide process adjustment in subsequent production batches. Feature-importance analysis also provides insights for process optimization, suggesting that reducing wafer-to-wafer variation in IV characteristics may yield disproportionately large benefits for the overall finished-product yield.

6. Conclusions

This study established an effective machine-learning framework for predicting the yield of CZT detector materials based on multi-wafer characterization data. By combining comprehensive electrical performance and spectral measurements with advanced machine-learning techniques, a predictive model was developed that can significantly optimize the quality-evaluation workflow. Through Bootstrap resampling of benchmark-wafer and evaluation-wafer data, the model can accurately predict the yield of test wafers even under conditions with only a limited number of ingot samples. By

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026 doi:10.20944/preprints202604.1691.v1

combining model-based yield prediction with pass/fail judgment based on empirical rules, the system can further improve the accuracy of ingot screening and thereby substantially reduce process cost.

This method not only provides practical utility for ingot-material screening but also offers insights into the factors most closely associated with high finished-product yield. Future work will focus on extending this method to other material systems and incorporating real-time process data for dynamic yield optimization.

Author Contributions: Conceptualization, [R.J./F.Y.]; methodology, [R.J.]; software, [R.J.]; validation, [R.J./ F.Y.]; formal analysis, [R.J.]; investigation, [R.J.]; resources, [F.Y./S.X./S.T.]; data curation, [F.Y.]; writing—original draft preparation, [R.J.]; writing—review and editing, [R.J./F.Y./W.J.]; visualization, [R.J.]; supervision, [W.J.]; project administration, [R.J./S.X./S.T.]; funding acquisition, [R.J./S.X.]. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFB3213203).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from the corresponding author upon reasonable request. Restrictions may apply due to industrial confidentiality.

Acknowledgments: The authors thank all colleagues involved in CZT ingot processing and wafer characterization for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

CZT	Cadmium Zinc Telluride
IV	Current–Voltage
FWHM	Full Width at Half Maximum
RF	Random Forest
LR	Linear Regression
MLR	Multiple Linear Regression
FNN	Feedforward Neural Network
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
MSE	Mean Squared Error

Reference

1. Fu, X.; Wang, F.B.; Zuo, X.R.; Wang, Z.J.; Wang, Q.R.; Wang, K.Q.; Xu, L.Y.; Xu, Y.D.; Guo, R.R.; Yu, H.; et al. Distinctive distribution of defects in CdZnTe: In ingots and their effects on the photoelectric properties. *Chinese Physics B* **2018**, *27*, 037302.
2. Roy, U.N.; Camarda, G.; Cui, Y.; Gul, R.; Hossain, A.; Yang, G.; Zazvorka, J.; Dedic, V.; Franc, J.; James, R. Role of selenium addition to CdZnTe matrix for room-temperature radiation detector applications. *Scientific reports* **2019**, *9*, 1620.
3. Ballester, M.; Kaspar, J.; Massanés, F.; Banerjee, S.; Vija, A.H.; Katsaggelos, A.K. Characterization of Crystal Properties and Defects in CdZnTe Radiation Detectors. *Crystals* **2024**, *14*, 935. https://doi.org/10.3390/cryst14110935.
4. Roy, U.N.; Camarda, G.S.; Cui, Y.; James, R.B. Optimization of selenium in CdZnTeSe quaternary compound for radiation detector applications. *Applied Physics Letters* **2021**, *118*.
5. Kirschenmann, S.; Bezak, M.; Bharthuar, S.; Brücken, E.; Golovleva, M.; Gädda, A.; Kalliokoski, M.; Karadzhinova-Ferrer, A.; Koponen, P.; Kramarenko, N.; et al. Quality assessment of cadmium telluride as a detector material for multispectral medical imaging. *Journal of Instrumentation* **2022**, *17*, C01070.
6. Amato, U.; Antoniadis, A.; De Feis, I.; Doinychko, A.; Gijbels, I.; La Magna, A.; Pagano, D.; Piccinini, F.; Selvan Suviseshamuthu, E.; Severgnini, C.; et al. Detecting Important Features and Predicting Yield from Defects Detected by SEM in Semiconductor Production. *Sensors* **2025**, *25*, 4218.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 April 2026
doi:10.20944/preprints202604.1691.v1

7. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.
8. Kim, H.; Kim, J.; Cho, S. Application of SMOTE for semiconductor manufacturing process with class imbalance. In Proceedings of the Proceedings of the Korean Operations Research and Management Science Conference. Korean Operations Research and Management Science Society, 2017, pp. 345–365.
9. Bai, J.; Wu, D.; Shelley, T.; Schubel, P.; Twine, D.; Russell, J.; Zeng, X.; Zhang, J. A Comprehensive Survey on Machine Learning Driven Material Defect Detection. *ACM Computing Surveys* **2025**, *57*, 1–36.
10. Khan, K.; Walker, D.M.H. A review of yield modelling techniques for semiconductor manufacturing. *International Journal of Production Research* **2024**, *62*, 6543–6562. https://doi.org/10.1080/00207543.2024.2351234.
11. Wang, S.; Chen, Y. Improved Yield Prediction and Failure Analysis in Semiconductor Manufacturing with XGBoost and Shapley Additive exPlanations Models. In Proceedings of the 2024 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA). IEEE, IEEE, 2024, p. 262. https://doi.org/10.1109/IPFA59546.2024.10638587.
12. Goldstein, H. *Multilevel statistical models*, 4th ed.; Wiley: Chichester, UK, 2011.
13. Steele, F. Multilevel models for longitudinal data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **2008**, *171*, 5–19. https://doi.org/10.1111/j.1467-985X.2007.00509.x.
14. Groß, J. *Linear regression*; Vol. 175, Springer Science & Business Media, 2003.
15. Marquardt, D.W.; Snee, R.D. Ridge regression in practice. *The American Statistician* **1975**, *29*, 3–20.
16. Ranstam, J.; Cook, J.A. LASSO regression. *Journal of British Surgery* **2018**, *105*, 1348–1348.
17. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
18. Chen, T. XGBoost: A Scalable Tree Boosting System. *Cornell University* **2016**.
19. LeCun, Y.; Touresky, D.; Hinton, G.; Sejnowski, T. A theoretical framework for back-propagation. In Proceedings of the Proceedings of the 1988 connectionist models summer school, 1988, Vol. 1, pp. 21–28.
20. Bebis, G.; Georgiopoulos, M. Feed-forward neural networks. *Ieee Potentials* **2002**, *13*, 27–31.
21. Krogh, A.; Hertz, J. A simple weight decay can improve generalization. *Advances in neural information processing systems* **1991**, *4*.
22. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **2014**, *15*, 1929–1958.
23. Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **1979**, *7*, 1–26. https://doi.org/10.1214/aos/1176344552.
24. Hesterberg, T. Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics* **2011**, *3*, 497–526.
25. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *The journal of machine learning research* **2012**, *13*, 281–305.
26. Snoek, J.; Larochelle, H.; Adams, R.P. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* **2012**, *25*.
27. Browne, M.W. Cross-validation methods. *Journal of mathematical psychology* **2000**, *44*, 108–132.
28. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.