Data-Efficient Conformalized Interval Prediction of Minimum Operating Voltage Capturing Process Variations

Yuxuan Yin
Electrical and Computer Engineering
University of California
Santa Barbara, CA, USA
y_yin@ucsb.edu

Rebecca Chen
Automotive Processing
NXP Semicondoctors
Austin, TX, USA
rebecca.chen_1@nxp.com

Chen He
Automotive Processing
NXP Semicondoctors
Austin, TX, USA
chen.he@nxp.com

Peng Li
Electrical and Computer Engineering
University of California
Santa Barbara, CA, USA
lip@ucsb.edu

ABSTRACT

Accurate minimum operating voltage (V_{min}) prediction is a critical element in manufacturing tests. Conventional methods lack coverage guarantees in interval predictions. Conformal Prediction (CP), a distribution-free machine learning approach, excels in providing rigorous coverage guarantees for interval predictions. However, standard CP predictors may fail due to a lack of knowledge of process variations. We address this challenge by providing principled conformalized interval prediction in the presence of process variations with high data efficiency, where the data from a few additional chips is utilized for calibration. We demonstrate the superiority of the proposed method on industrial 16nm chip data.

1 INTRODUCTION

The assessment of the minimum operating voltage (V_{min}) stands as an important test for silicon manufacturers aiming to enhance product reliability and yield by identifying and eliminating outliers and potentially defective chips [5]. Nevertheless, the precise determination of V_{min} necessitates the execution of a testing scheme multiple times, spanning a broad range of operating voltages. This limitation hampers its practical applicability due to the associated undesirable temporal and financial costs.

In light of the emerging success of contemporary Machine-Learning (ML) methods across diverse scientific domains, there is a growing focus on developing an ML-based chip performance predictor to expedite the testing workflow [1, 7, 12, 16]. The key

(a) Lot-to-lot variations
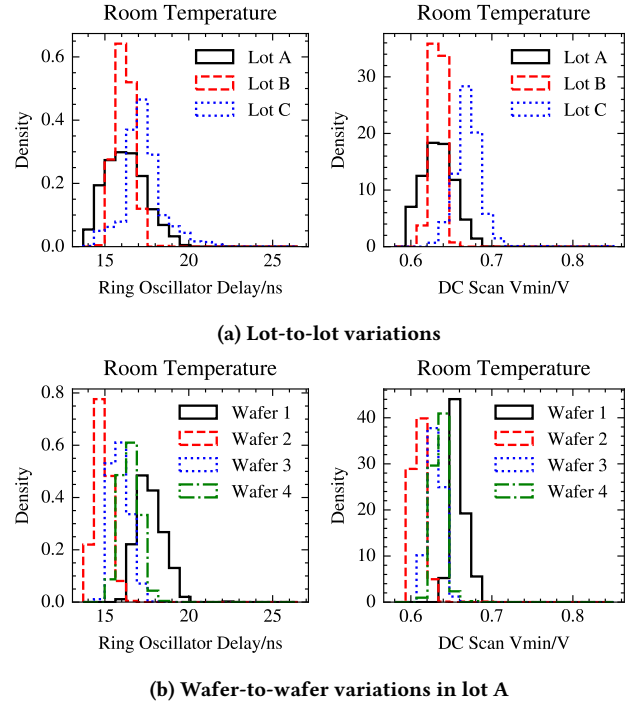

(b) Wafer-to-wafer variations in lot A

Figure 1: Process variations of example 16nm chips

idea is to leverage informative, cost-effective testing features, such as on-chip Ring Oscillator (RO) delay and IDDQ leakage current, to predict V_{min}, whose direct measurement is expensive.

Nevertheless, the accuracy of V_{min} predicted by an ML technique may be compromised by measurement noise, testing limitations, and model capacity constraints, falling short of the high reliability demanded by the semiconductor industry. Adding to the challenge, the presence of process variations between training and testing chips further impedes the accuracy of the V_{min} predictor. As illustrated in Figure 1, both lot-to-lot and wafer-to-wafer variations manifest in the input features and target V_{min} values across a 16nm chip dataset, consisting of multiple lots with multiple wafers per
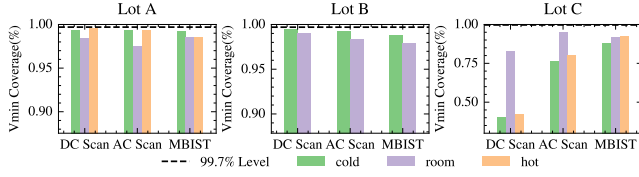
Figure 2: Coverage rate of 6-sigma intervals of multiple lots predicted by linear regression trained on lot A

lot. Consequently, it is risky to directly employ a pre-trained V_{min} predictor on chips from different lots/wafers. To this end, it is desirable to develop a robust interval predictor that ensures ❶ a rigorous coverage for V_{min} prediction under ❷ process variations, while also exhibiting ❸ high data efficiency.

Traditional interval prediction methods [8], such as ones based on utilizing the 6-sigma of the error distribution, focus on achieving a 99.7% coverage rate while making the strong Gaussian assumption on the V_{min} predictor's error. However, the underlying assumption for error distribution may be misleading, especially when the training data is insufficient to fit an accurate V_{min} predictor. Furthermore, process variations may exacerbate the deviation from any derived theoretical coverage. In Figure 2, we illustrate the empirical coverage rate of the 6-sigma interval on hundreds of chips from all of the 3 lots, whereas the V_{min} predictor is a linear regressor trained on thousands of chips from lot A. Additional details regarding experimental settings are provided in Section 4.1. Notably, the 6-sigma method not only falls short of the theoretical coverage rate on lot A, indicating an incorrect assumption about the error distribution, but also performs poorly under the influence of process variations, particularly evident in predictions made on lot C.

Conformal Prediction (CP) [11] has emerged as a promising distribution-free Uncertainty Quantification (UQ) method for constructing intervals based on any point predictor while offering a nonasymptotic coverage guarantee. CP leverages held-out calibration data, separated from the training dataset, to assess the uncertainty of a fitted regression model for new testing data. While adopting CP for reliable V_{min} interval prediction is a plausible implementation to address coverage guarantee, this attribute is no longer held when utilizing the CP predictor to assess chips manufactured under another process condition.

To this end, we propose a novel distribution-free V_{min} interval prediction framework, named Calibrated Weighed Conformalized Quantile Regression (CWCQR), with a ❶ theoretical coverage guarantee in addressing both ❷ process variations and ❸ data efficiency. Our approach systematically calibrates Conformalized Quantile Regression (CQR) [10] to satisfy a designed coverage rate under process variations. Our primary contributions are:

• We incorporate Weighted Conformal Prediction (WCP) [14] into CQR to handle the distribution shift of input features under process variations without requiring any new chip data.

• We propose a calibration method to transform the prediction intervals of a pre-trained CP predictor to accommodate shifts in V_{min} mean value and noise under the new processing condition, leveraging only a small number of newly measured chips.

• We empirically demonstrate that our method outperforms several existing interval prediction baselines on an industrial dataset.

2 BACKGROUND

We first describe the general training flow of ML-based V_{min} predictors, then we summarize prevalent interval prediction methods.

2.1 Point Estimation

V_{min} point prediction can be viewed as fitting an ML-based predictor to minimize a loss function on training chips, with low-cost test features as model input, and V_{min} as output. We denote a training dataset containing N chips as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, whose D test features are aggregated as a vector $\mathbf{x} \in \mathbb{R}^D$, and true V_{min} is a scalar $y \in \mathbb{R}$. We can optimize parameters $\boldsymbol{\theta}$ of a predictor $g(\cdot; \boldsymbol{\theta}): \mathbb{R}^D \to \mathbb{R}$ by minimizing a loss function $\mathcal{L}$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(g(\mathbf{x}_i; \boldsymbol{\theta}), y_i). \tag{1}$$

where $\hat{y} = g(\mathbf{x}; \boldsymbol{\theta})$ is the model prediction.

For instance, we can apply the Mean Squared Error (MSE) loss $\mathcal{L}_{mse}$ to train a point predictor with small bias and variance.

2.2 Interval Prediction

V_{min} point estimation is not the final step in the testing process. Indeed, a V_{min} interval needs to be constructed to differentiate between normal parts and outliers. Silicon industry engineers practically apply a 6-sigma width interval to a point prediction. For a chip with input features $\mathbf{x}$, its 6-sigma interval $C(\mathbf{x})$ is

$$C(\mathbf{x}) = [g(\mathbf{x}; \boldsymbol{\theta}) - 3\sigma, g(\mathbf{x}; \boldsymbol{\theta}) + 3\sigma], \tag{2}$$

where σ is the standard deviation of g's prediction error on the training dataset $\mathcal{D}$.

Despite plausible bounds for chips constructed by the 6-sigma approach, this method is unable to provide a theoretical coverage rate for its prediction of testing data. Suppose we have a testing chip $(\mathbf{x}_{test}, y_{test})$, we want to know the empirical miscoverage rate $\alpha \in [0, 1]$ for the 6-sigma interval in Equation (2):

$$\alpha = 1 - \mathbb{P}\{y_{test} \in C(\mathbf{x}_{test})\}. \tag{3}$$

Unfortunately, the estimation of α relies on strong assumptions on the model prediction's errors, limiting the reliability and efficacy of this heuristic interval prediction approach.

2.3 Conformal Prediction (CP)

CP [11] is a distribution-free UQ method to provide a rigorous nonasymptotic coverage guarantee for testing data. It is an additional post-training strategy for a point predictor to construct reliable intervals. The prerequisite assumption of CP is that training data and testing data are *exchangeable*. The exchangeability is a weaker condition than the i.i.d. property, i.e., $\mathcal{D}$ and $(\mathbf{x}_{test}, y_{test})$ are sampled from an identical distribution. In our chip performance prediction case, the assumption is supposed to be held, as long as both sets of chips come from the same process.

CP aims to construct an interval $C(\cdot)$ with a designed coverage rate $1 - \alpha$ for a point predictor g on the testing chip:

$$C(\mathbf{x}) = \{y : s(\mathbf{x}, y) \leq \hat{q}\}, \tag{4}$$

$$s.t. \quad \mathbb{P}\{y_{test} \in C(\mathbf{x}_{test})\} \geq 1 - \alpha, \tag{5}$$

where $\hat{q} \in [0, +\infty)$ is a positive value to be determined, $s(\mathbf{x}, y) \rightarrow \mathbb{R}$ is a score function to evaluate the conformity.

We describe how to compute $\hat{q}$ in CP, specifically, in *split* CP [1]. Split CP divides the training dataset $\mathcal{D}$ into two disjoint subsets, where one is for training the model, and the other is for computing $\hat{q}$. The idea is to calibrate the model's heuristic uncertainty by testing the conformity of some unseen data, sampled from the distribution of testing data $(\mathbf{x}_{test}, y_{test})$. The conformity is evaluated by a score function $s(\mathbf{x}, y) \rightarrow \mathbb{R}$, whose larger value means higher discrepancy.

We dive into the detailed procedures of CP. The two disjoint subsets partitioned from $\mathcal{D}$ are a new training dataset $\mathcal{D}_{train}$ and calibration dataset $\mathcal{D}_{cali}$. A point predictor g is trained on the new training dataset $\mathcal{D}_{train}$ to minimize Equation (1).

Without loss of generality, we suppose the first $|\mathcal{D}_{cali}|$ items form the calibration set. Next, we choose the absolute residual as the conformal score, and compute the conformity of the calibration dataset $\mathcal{D}_{cali}$:

$$s_i = s(\mathbf{x}_i, y_i) := |g(\mathbf{x}_i; \boldsymbol{\theta}) - y_i|, \quad i = 1, \cdots, |\mathcal{D}_{cali}|. \tag{6}$$

Finally, $\hat{q}$ is set to the $\lceil (1 - \alpha)(1 + |\mathcal{D}_{cali}|) \rceil$th-smallest score:

$$\hat{q} = \text{Quantile}\left(s_1, \cdots, s_{|\mathcal{D}_{cali}|}; \frac{\lceil (1 - \alpha)(1 + |\mathcal{D}_{cali}|) \rceil}{|\mathcal{D}_{cali}|}\right). \tag{7}$$

It is proved in [15] that $\hat{q}$ in Equation (7) satisfies the coverage rate $1 - \alpha$ for the interval in Equation (4), which is equivalent to

$$C(\mathbf{x}) = [g(\mathbf{x}; \boldsymbol{\theta}) - \hat{q}, g(\mathbf{x}; \boldsymbol{\theta}) + \hat{q}]. \tag{8}$$

Despite CP provides distribution-free prediction intervals with a theoretical coverage guarantee, it still has an apparent limitation: the length of intervals is fixed to $2\hat{q}$, regardless of the input features $\mathbf{x}_{test}$ of the testing chip. This characteristic potentially leads to underkill of outliers and overkill of normal chips. Moreover, it fails to address process variations, restricting its effectiveness in industrial practice.

2.4 Conformalized Quantile Regression (CQR)

CQR [10] introduces Quantile Regression (QR) [6] to CP for the first time to enable adaption to heteroscedasticity, meanwhile maintaining the equivalent nonasymptotic coverage guarantee. The framework of CQR is the same as that of CP, except for 2 differences:

• CQR adopts two quantile regressors to provide heuristic interval prediction;

• CQR proposes a new score function to assess the conformity.

Given a designed coverage rate $1 - \alpha$, CQR trains two quantile predictors, $g_{\alpha/2}(\cdot; \boldsymbol{\theta}_{lo}) : \mathbb{R}^D \rightarrow \mathbb{R}$ and $g_{1-\alpha/2}(\cdot; \boldsymbol{\theta}_{hi}) : \mathbb{R}^D \rightarrow \mathbb{R}$, to regress the lower and the upper bound of a $(\alpha/2)$-th to $(1 - \alpha/2)$-th quantile interval, respectively. the quantile loss $\mathcal{L}_\gamma$ instead of the MSE loss $\mathcal{L}_{mse}$ is set for model training:

$$\mathcal{L}_\gamma(\hat{y}, y) := \max\{\gamma(y - \hat{y}), (1 - \gamma)(\hat{y} - y)\}, \tag{9}$$

---

[1]CP has two primary variants: *full* CP and *split* CP. Full CP is not applicable for regression problems when y can take infinite numbers of values, since it requires infinite times of model training. Thereby, we introduce split CP herein.

where $\gamma \in [0, 1]$ is the designed quantile.

In CQR, the conformal score of i-th item of calibration dataset $\mathcal{D}_{cali}$ is defined as

$$s_i := \max\{g(\mathbf{x}_i; \boldsymbol{\theta}_{lo}) - y_i, \quad y_i - g(\mathbf{x}_i; \boldsymbol{\theta}_{hi})\}. \tag{10}$$

CQR employs a scalar $\hat{q}$ for the prediction interval, where $\hat{q}$ is also the the $\lceil (1 - \alpha)(1 + |\mathcal{D}_{cali}|) \rceil / |\mathcal{D}_{cali}|$-th quantile of conformal scores, computed in Equation (7). The formulation of prediction interval in Equation (4) can be written as

$$C(\mathbf{x}) = [g(\mathbf{x}; \boldsymbol{\theta}_{lo}) - \hat{q}, \quad g(\mathbf{x}; \boldsymbol{\theta}_{hi}) + \hat{q}]. \tag{11}$$

As shown in [10], the aforementioned interval satisfies $1 - \alpha$ coverage rate in Equation (5).

Even though CQR gives a prediction interval for the testing chip $(\mathbf{x}_{test}, y_{test})$ with a nonasymptotic coverage guarantee, it assumes that $(\mathbf{x}_{test}, y_{test})$ and $(\mathbf{x}_i, y_i)$ follows the same distribution, where $(\mathbf{x}_i, y_i) \in \mathcal{D}$. Due to the existence of process variations as shown in Figure 1, the assumption may be violated when $(\mathbf{x}_{test}, y_{test})$ is picked from a new lot or a new wafer, thus deteriorating the coverage rate of CQR.

3 PROPOSED METHOD

In order to stress process variations, we propose CWCQR, satisfying a nonasymptotic coverage guarantee with high data efficiency. We introduce weighted CP [14] to CQR to handle the distribution shift of input features between training data and testing data. Moreover, we propose an additional calibration scheme to translate and stretch prediction intervals such that the designed coverage rate is guaranteed while preserving invaluable model parameters learned from a large training dataset $\mathcal{D}$. CWCQR merits data efficiency in ensuring reliable prediction intervals.

3.1 Weighted CQR (WCQR)

Weighted CP [14] is capable of a specific kind of process variation, named *covariate shift* [13], which means training features' distribution p and testing features' distribution p_{test} are different, whereas the conditional distribution $y|\mathbf{x}$ stays fixed. This relationship can be formulated as:

$$y = f(\mathbf{x}) + \delta(\mathbf{x}), \tag{12}$$

where $f : \mathbb{R}^D \rightarrow \mathbb{R}$ is an agnostic real-world physical model, $\delta : \mathbb{R}^D \rightarrow \mathbb{R}$ is an unknown zero-mean noise, dependent on $\mathbf{x}$.

Weighted CP introduces a weighting scheme for conformal scores, whose weights depend on features of the testing data. The weight of i-th conformal score s_i is

$$w_i(\mathbf{x}_{test}) := \frac{r(\mathbf{x}_i)}{r(\mathbf{x}_{test}) + \sum_{j=1}^{|\mathcal{D}_{cali}|} r(\mathbf{x}_j)}, \tag{13}$$

where $r : \mathbb{R}^D \rightarrow [0, +\infty)$ is the likelihood ratio of two feature distributions: $r(\mathbf{x}) = p_{test}(\mathbf{x})/p(\mathbf{x})$.

Intuitively, the weight is larger when the features are more likely to be sampled from the test distribution, and vice versa.

Next, we describe how weighted CP computes $\hat{q}(\mathbf{x}_{test})$, which is dependent on $\mathbf{x}_{test}$. Without loss of generality, suppose $\mathcal{D}_{cali}$ is sorted to an ascending order of the conformal score. Then, we have

$$\hat{q}(\mathbf{x}_{test}) = \inf\left\{s_j : \sum_{i=1}^{j} w_i(\mathbf{x}_{test}) \geq 1 - \alpha\right\}. \tag{14}$$

Table 1: Description of a 16nm industrial dataset

Attributes	Lot A	Lot B	Lot C
# Dies	3754	731	754
# Parametric features	354	354	354
# DC Scan V_{min}	13	13	13
# AC Scan V_{min}	13	13	13
# MBIST V_{min}	9	9	9
Test temperature/°C	-40, 25, 125	-40, 25	-40, 25, 125

such that Equation (5) is satisfied where

$$C(\mathbf{x}_{test}) = \{y : s(\mathbf{x}_{test}, y) \le \hat{q}(\mathbf{x}_{test})\}. \tag{15}$$

While weighed CP is a general framework regardless of the selection of the score function, its combination with CQR has not been explored yet, to the best of our knowledge. Moreover, weighed CP provides a plausible solution for covariate shift in process variations. To this end, we propose Weighted Conformalized Quantile Regression (WCQR), introducing the weighed scheme to CQR.

In WCQR, the score function is defined in Equation (10), and the prediction interval is

$$C(\mathbf{x}_{test}) = \left[g(\mathbf{x}_{test}; \boldsymbol{\theta}_{lo}) - \hat{q}(\mathbf{x}_{test}), \quad g(\mathbf{x}_{test}; \boldsymbol{\theta}_{hi}) + \hat{q}(\mathbf{x}_{test})\right], \tag{16}$$

where $\hat{q}(\mathbf{x}_{test})$ is presented in Equation (14). Under the weighted CP's framework [14], our prediction interval has at least $1 - \alpha$ coverage rate for y_{test}.

3.2 Calibrated WCQR (CWCQR)

The proposed WCQR provides interval prediction with a nonasymptotic coverage guarantee. Note that such property is only held under the assumption of covariate shift: $y|\mathbf{x}$ is the same for training data and testing data. This assumption, however, is sometimes too strong in our V_{min} prediction scenarios, where some process variations do not essentially follow it. A more general and realistic assumption is that $y_{test}|\mathbf{x}_{test}$ has a constant shift with a changed noise intensity, diverged from $y|\mathbf{x}$. This assumption on $y_{test}|\mathbf{x}_{test}$ can be formulated as:

$$y_{test} = f(\mathbf{x}_{test}) + b' + \delta'(\mathbf{x}_{test}), \tag{17}$$

where bias $b' \in \mathbb{R}$ is a real number, noise δ' is sampled from a zero-mean distribution dependent on $\mathbf{x}_{test}$.

We propose CWCQR to handle the aforementioned type of process variations. Its prediction interval for $\mathbf{x}_{test}$ is constructed as

$$C(\mathbf{x}_{test}) = [y_{low}, y_{high}],$$

$$\text{where} \quad y_{low} = g(\mathbf{x}_{test}; \boldsymbol{\theta}_{lo}) - \hat{q}(\mathbf{x}_{test}) + \hat{b}_{cali} - \hat{q}_{cali}, \tag{18}$$

$$y_{high} = g(\mathbf{x}_{test}; \boldsymbol{\theta}_{hi}) + \hat{q}(\mathbf{x}_{test}) + \hat{b}_{cali} + \hat{q}_{cali}.$$

Herein, two scalars $\hat{b}_{cali}$ and $\hat{q}_{cali}$ are determined to address the bias b' and the noise δ' in ensuring a $1 - \alpha$ coverage rate for y_{test}.

CWCQR employs a small additional calibration dataset $\mathcal{D}'_{cali}$ to compute $\hat{b}_{cali}$ and $\hat{q}_{cali}$, where $\mathcal{D}'_{cali}$ and $(\mathbf{x}_{test}, y_{test})$ are under the same process. For convenience, we suppose the index for $\mathcal{D}'_{cali}$ takes value from $N + 1$ to $N + M$.

Table 2: V_{min} prediction error of CatBoost

Lot_{te}	Temperature	Lot_{tr}	V_{min} Prediction RMSE (mV)		
			DC Scan	AC Scan	MBIST
A	cold	A	5.52	7.17	8.84
	room		22.89	29.54	22.67
	hot		6.85	8.35	15.34
B	cold	A	6.14	8.89	10.67
		B	5.09	7.13	8.16
	room	A	20.55	25.82	25.37
		B	16.66	23.51	20.52
C	cold	A	29.92	25.23	23.54
		C	9.36	9.59	10.79
	room	A	31.45	25.96	29.31
		C	15.58	13.97	11.89
	hot	A	26.07	25.74	26.49
		C	9.85	10.64	14.46

For the estimation of V_{min} shift, we translate prediction intervals $C(\mathbf{x})$ to $C(\mathbf{x}) + \hat{b}_{cali}$ to cover $\mathcal{D}'_{cali}$ as much as possible:

$$\hat{b}_{cali} = \arg\max_b |\{i : y_i \in C(\mathbf{x}_i) + b, i = N + 1, \cdots, N + M\}|. \tag{19}$$

In tackling the discrepant noise under process variations, we stretch prediction intervals to satisfy the coverage rate $1-\alpha$. We adopt ideas from CQR to compute the conformity of $\mathcal{D}'_{cali}$, whose conformal score is

$$s_i := \max \left\{g(\mathbf{x}_i; \boldsymbol{\theta}_{lo}) - (y_i - \hat{b}_{cali}), \quad y_i - \hat{b}_{cali} - g(\mathbf{x}_i; \boldsymbol{\theta}_{hi})\right\}. \tag{20}$$

Next, we select $\hat{q}_{cali}$ as

$$\hat{q}_{cali} = \text{Quantile}\left(s_{N+1}, \cdots, s_{N+M}; \frac{\lceil(1 - \alpha)(1 + M)\rceil}{M}\right). \tag{21}$$

4 EXPERIMENTS

We conduct experiments to demonstrate the efficacy of our approach for addressing process variations on thousands of 16nm automotive chips. We aim to illustrate: 1) the fatal impact of process variations in standard V_{min} point predictors, and 2) CWCQR's superior efficacy and data efficiency in addressing process variations. In specific, we consider *lot-to-lot* variations in our experiments.

As shown in Table 1, chips are sampled from 3 different lots: 3754 from lot A, 731 from lot B, and 754 from lot C. For each chip, we employ the same test flow to collect 354 parametric features and V_{min}, used for input data and output targets of predictors, respectively. 3 types of V_{min} are measured: 13 patterns of DC Scan V_{min}, 30 patterns of AC Scan V_{min}, and 21 patterns of MBIST V_{min}. We describe how the input features and the output V_{min} are collected.

In the testing flow, We first test all patterns of V_{min}, and then perform the parametric tests. Both phases are done at the same specific temperature: -45℃ (cold), 25℃ (room), or 125℃ (hot). Both V_{min} and parametric tests are performed on an Automatic Test Equipment (ATE) tester.

Table 3: Comparison of regressors in lot A at cold temperature

V_{min} predictor	V_{min} Prediction RMSE (mV)		
	DC Scan	AC Scan	MBIST
Linear Regression	10.65	10.99	12.88
XGBoost	5.63	8.23	9.25
CatBoost	**5.52**	**7.17**	**8.84**

4.1 Impact of Lot-to-Lot Variations in Vmin Point Estimation

We empirically illustrate the influence of lot-to-lot variations for V_{min} predictors.

Experimental Settings. We train A V_{min} predictor solely on lot A, and then directly employ it to predict chips from lot B and lot C. In order to evaluate how its accuracy deviates when no process variations exist, we further report the performance of a predictor, whose training and testing data are from the same lot. In this case, 4-fold cross-validation is applied to split the full lot into two subsets. The V_{min} point predictor is CatBoost [9], which performs well in our dataset.

Results. The V_{min} prediction error of CatBoost is shown in Table 2, where Lot$_{te}$ and Lot$_{tr}$ represent the testing and training lot, respectively. The lot-to-lot variation is relatively small between lots A and B, and is large between lots A and C, across all testing temperatures and V_{min} test patterns: the increase of RMSE is up to 5mV, 25% in the first case, and 21mV, 220% in the second case. This outcome remains consistent with (a) in Figure 1, where the discrepancy between lots A and C is much larger than that between lots A and B, in terms of both input features and target V_{min}.

Our experiment indicates that The magnitude of lot-to-lot variations varies in a large range. A consequence is that directly adopting a predictor for new chips is risky, since the discrepancy between training and new chips is agnostic to us, advancing to any evaluation of new data. To this end, an additional calibration stage is essential for a pre-trained V_{min} predictor to achieve high reliability.

Remarks. We demonstrate the reason for applying CatBoost in V_{min} prediction. CatBoost is compared with two baselines: linear regression and XGBoost [2]. For linear regression, 5 features are selected from a total of 354 parametric data by Correlation Feature Selection (CFS) [4] with the Pearson correlation [3] for the model's input. A 4-fold cross-validation is performed in lot A at cold temperature to report results. As shown in Table 3, CatBoost achieves superior performance in V_{min} prediction average across test patterns.

4.2 Efficacy of CWCQR

Our experiments compare CWCQR to vanilla CQR on our industrial dataset to verify its effectiveness in addressing lot-to-lot variations in V_{min} interval prediction.

Experimental Settings. For both methods, we employ the same configurations for a fair comparison. Since CWCQR basically calibrates a base CQR model, we first introduce how to train a CQR, then describe how our method works.
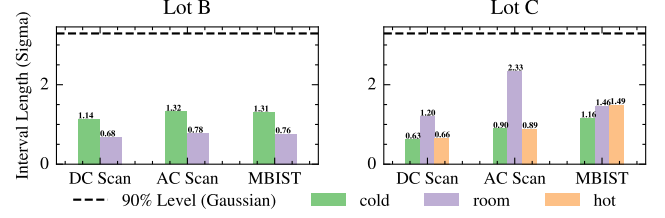
Figure 3: Average sigma of CWCQR with a 90% Coverage rate

For CQR, the training dataset $\mathcal{D}$ is all of 3754 chips from lot A, and it is further separated into two subsets: 1) $\mathcal{D}_{train}$ for containing 75% chips for training quantile regressors, and 2) $\mathcal{D}_{cali}$, including the rest 25% ones for calibration. We set the miscoverage α to 0.1, and train two CatBoost quantile regressors whose objectives are 5% and 95% V_{min}, respectively.

For CWCQR, two additional steps are proposed: a weighting scheme and a further calibration. In the first stage, we follow [14] to adopt a logestic classifier to estimate the likelihood ratio. In avoiding the overfitting problem for this classifier, CFS is employed to select 5 features with the top-5 highest linear correlations to V_{min}. In the second stage, 50 evaluated from the testing lot are utilized for further calibration to compute $\hat{b}_{cali}$ and $\hat{q}_{cali}$.

One unfair setting is that CWCQR essentially requires new calibration data. To address this concern, we also calibrate the CQR model on these data: the model is viewed as a heuristic interval predictor just like QR, and a CQR framework is employed to satisfy a 90% coverage rate.

Results. The average length and coverage rate of V_{min} prediction intervals are listed in Table 4. Although CQR satisfies the designed coverage in lot A, its reliability is not held in lot B and C. In the worst case, it only covers less than 12% true DC Scan V_{min} of chips from lot C, tested at the hot temperature. Despite that $\mathcal{D}'_{cali}$ successfully calibrates CQR to achieve a 90% coverage rate, the average length of its prediction intervals is much longer than that of CWCQR for most V_{min} test patterns, underscoring the superior performance of our approach. Compared with CQR trained on the 75% new data (∼550 chips), CWCQR is much more data-efficient: it requires measurements of 50 new chips for calibration. We save 91% testing time to provide a reliable interval prediction with a 90% coverage for V_{min}, scarifying an increased interval length within 0.7-sigma for DC Scan V_{min}, 0.6-sigma for AC Scan V_{min}, and 1-sigma for MBIST V_{min}, where sigma is the V_{min} point predictor's residual under process variations.

In addition, we demonstrate that the distribution-free CWCQR outperforms the sigma method to satisfy a 90% coverage demand. We report the prediction interval length quantified by the sigma average across all test patterns under lot-to-lot variations. As shown in Figure 3, CWCQR achieves much smaller intervals than the sigma method, indicating that CWCQR's efficiency and robustness under process variations.

Table 4: Length and coverage of V_{min} prediction intervals provided by CatBoost for a designed 90% coverage rate

Lot$_{te}$	Temperature	Lot$_{tr}$	Method	DC Scan		AC Scan		MBIST	
				Len. (mV)	Cov. (%)	Len. (mV)	Cov. (%)	Len. (mV)	Cov. (%)
B	cold	B	CQR	19.55	90.98±1.45	26.52	90.60±1.18	31.18	90.18±1.36
		A	CQR	17.25	86.87±3.36	24.44	87.93±3.64	28.26	79.05±11.86
			CQR with Addi. Cali.	21.39	93.02±3.78	30.10	92.57±6.33	35.54	90.54±3.87
			CWCQR	**20.63**	**90.64±4.94**	**27.57**	**90.26±6.21**	**36.31**	**90.93±4.39**
	room	B	CQR	23.46	91.58±1.20	32.74	90.45±1.31	38.30	90.74±0.88
		A	CQR	34.38	92.24±2.07	49.37	90.83±2.62	42.17	75.47±21.72
			CQR with Addi. Cali.	35.27	92.29±4.47	53.95	93.24±2.51	52.40	90.55±4.29
			CWCQR	**31.94**	**89.38±4.40**	**46.09**	**91.11±4.49**	**42.31**	**87.02±6.17**
C	cold	C	CQR	28.13	90.79±0.67	34.55	91.46±0.81	42.05	91.59±1.21
		A	CQR	17.00	12.57±15.69	23.09	23.94±9.41	27.61	45.48±20.28
			CQR with Addi. Cali.	77.21	91.51±7.16	71.31	90.47±3.88	77.39	91.03±3.37
			CWCQR	**36.73**	**91.38±4.42**	**39.02**	**90.25±3.68**	**56.43**	**90.05±6.17**
	room	C	CQR	30.06	91.90±1.03	37.03	91.77±0.91	46.30	91.91±0.70
		A	CQR	41.35	44.03±38.57	57.40	72.93±27.77	42.20	58.14±25.67
			CQR with Addi. Cali.	85.38	92.56±4.63	71.34	91.60±5.20	91.97	93.47±3.64
			CWCQR	**40.77**	**89.65±3.91**	**51.25**	**86.81±4.21**	**70.48**	**89.74±4.80**
	hot	C	CQR	29.75	90.13±1.66	42.70	91.70±2.20	55.05	91.82±2.48
		A	CQR	23.90	11.88±7.80	33.16	25.50±12.25	50.99	50.06±25.34
			CQR with Addi. Cali.	78.16	92.65±2.44	93.27	92.46±3.42	91.00	92.26±4.21
			CWCQR	**41.10**	**91.63±4.16**	**55.45**	**92.85±4.66**	**70.92**	**90.69±6.09**

5 CONCLUSION

We propose a novel distribution-free V_{min} interval prediction method, CWCQR, with a theoretical nonasymptotic coverage guarantee. A systematical calibration scheme is introduced to adjust a base CQR model to satisfy the designed coverage rate. Leveraging informative information from training lots, CWCQR achieves superior data efficiency and effectiveness. Although lot-to-lot variations are considered in our experiment for demonstration, we believe our method is capable of other kinds of process variations in silicon manufacturing flows. In fact, CWCQR is a new data-efficient conformalized interval prediction framework to calibrate any pre-trained interval predictor to ensure high reliability in addressing distribution shifts between training and testing data.

ACKNOWLEDGMENTS

REFERENCES

[1] Janine Chen, Jing Zeng, Li-C. Wang, Michael Mateja, and Jeff Rearick. 2010. Predicting multi-core system Fmax by data-learning methodology. In *Proceedings of 2010 International Symposium on VLSI Design, Automation and Test*. 220–223.
[2] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
[3] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.
[4] Mark A Hall. 1999. *Correlation-based feature selection for machine learning*. Ph. D. Dissertation. The University of Waikato.
[5] Chen He and Yanyao Yu. 2020. Wafer Level Stress: Enabling Zero Defect Quality for Automotive Microcontrollers without Package Burn-In. In *2020 IEEE International Test Conference (ITC)*. 1–10. https://doi.org/10.1109/ITC44778.2020.9325213
[6] Roger Koenker and Gilbert Bassett Jr. 1978. Regression quantiles. *Econometrica: journal of the Econometric Society* (1978), 33–50.
[7] Yen-Ting Kuo, Wei-Chen Lin, Chun Chen, Chao-Ho Hsieh, James Chien-Mo Li, Eric Jia-Wei Fang, and Sung S.-Y. Hsueh. 2021. Minimum Operating Voltage Prediction in Production Test Using Accumulative Learning. In *2021 IEEE International Test Conference (ITC)*. 47–52. https://doi.org/10.1109/ITC50571.2021.00012
[8] David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.
[9] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31.
[10] Yaniv Romano, Evan Patterson, and Emmanuel Candes. 2019. Conformalized quantile regression. *Advances in neural information processing systems* 32 (2019).
[11] Glenn Shafer and Vladimir Vovk. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research* 9, 3 (2008).
[12] Qihang Shi, Xiaoxiao Wang, LeRoy Winemberg, and Mark M Tehranipoor. 2016. On-chip sensor selection for effective speed-binning. *Analog Integrated Circuits and Signal Processing* 88 (2016), 369–382.
[13] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90, 2 (Oct. 2000), 227–244.
[14] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.
[15] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Vol. 29. Springer.
[16] Yuxuan Yin, Rebecca Chen, Chen He, and Peng Li. 2023. Domain-Specific Machine Learning Based Minimum Operating Voltage Prediction Using On-Chip Monitor Data. In *2023 IEEE International Test Conference (ITC)*. 99–104.