Reliable Interval Prediction of Minimum Operating Voltage Based on On-chip Monitors via Conformalized Quantile Regression

Yuxuan Yin
Dept. of ECE
University of California
Santa Barbara, USA
y_yin@ucsb.edu

Xiaoxiao Wang
NXP Semiconductors
Austin, USA
Xiaoxia3@andrew.cmu.edu

Rebecca Chen
NXP Semiconductors
Austin, USA
rebecca.chen_1@nxp.com

Chen He
NXP Semiconductors
Austin, USA
chen.he@nxp.com

Peng Li
Dept. of ECE
University of California
Santa Barbara, USA
lip@ucsb.edu

Abstract—**Predicting the minimum operating voltage (V_{min}) of chips is one of the important techniques for improving the manufacturing testing flow, as well as ensuring the long-term reliability and safety of in-field systems. Current V_{min} prediction methods often provide only point estimates, necessitating additional techniques for constructing prediction confidence intervals to cover uncertainties caused by different sources of variations. While some existing techniques offer region predictions, but they rely on certain distributional assumptions and/or provide no coverage guarantees. In response to these limitations, we propose a novel distribution-free V_{min} interval estimation methodology possessing a theoretical guarantee of coverage. Our approach leverages conformalized quantile regression and on-chip monitors to generate reliable prediction intervals. We demonstrate the effectiveness of the proposed method on an industrial 5nm automotive chip dataset. Moreover, we show that the use of on-chip monitors can reduce the interval length significantly for V_{min} prediction.**

Index Terms—**chip performance prediction, on-chip monitors, conformal prediction, quantile regression**

I. INTRODUCTION

Measurement of the minimum operating voltage (V_{min}) is one of the important testing procedures to determine chip performance. It facilitates the detection of inferior products, the conservation of power consumption, and the indication of potential early life failures. As technology nodes keep scaling, V_{min} tests via structural test patterns (e.g., SCAN) become more and more crucial and necessary to screen out tiny flaws and defects [1] inside chips.

Conventional V_{min} measurements involve testing chips at a high operating voltage and decreasing step by step until they fail, which is time-consuming. Moreover, such a strategy is exclusively applicable in the manufacturing test process, but not in-field systems. To this end, researchers propose to build machine learning based V_{min} predictors utilizing low-cost features, such as parametric testing data from the production test flow and on-chip monitor data for the in-field prediction [2]–[5]. Many regression models have been explored recently, including linear regression [4], Gaussian Process (GP) [3], and Neural Network (NN) [5]. For instance, Chen demonstrated a low-cost approach to predict the system F_{max} (the maximum operating frequency) using the structural F_{max} of flip flops [3] via a GP model, whose kernel hyperparameter length scales are used as indicators of the significance of features. Yin adopted a constrained NN to capture the monotonicity between RO delay and V_{min} degradation [5]. Although these methods provide promising point estimation for V_{min}, additional techniques are still required to construct prediction intervals to ensure high coverage of true V_{min} to account for the uncertainties due to variations of process, voltage, temperature, operating frequency, application mode, etc.

Uncertainty Quantification (UQ) for machine learning provides the model's confidence interval. Commonly employed UQ methods include 1) Bayesian approaches such as GP [6] and Bayesian neural networks [7], 2) neural networks ensemble [8], and 3) Quantile Regression (QR) [9]. While these methods excel at estimating uncertainty within the training data distribution, their prediction intervals often lack a reliable coverage guarantee for new testing data. Consequently, none of these approaches fully meet the stringent demands of the silicon industry for generating robust V_{min} intervals to ensure high reliability.

Conformal Prediction (CP) [10] emerges as a promising distribution-free UQ method for constructing intervals based on any point predictor while offering a nonasymptotic coverage guarantee. CP leverages a calibration dataset to assess the uncertainty associated with a fitted regression model by analyzing its prediction residuals. However, vanilla CP exhibits limitations as a V_{min} region predictor, as it constructs constant intervals for all testing samples, potentially leading to excessive margins

Table I: Comparison of uncertainty quantification methods

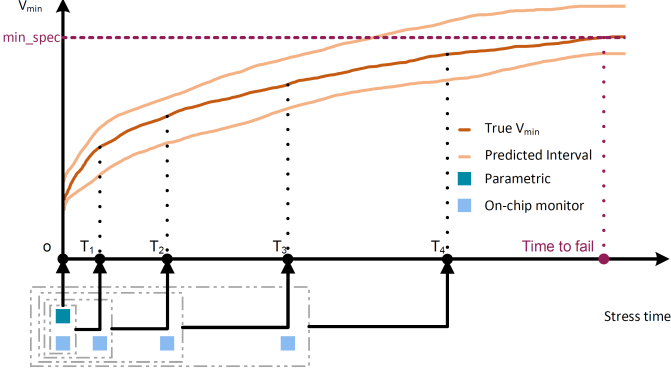| Property	Bayesian	Ensemble	QR	CP	CQR
Distribution-free	✗	✓	✓	✓	✓
Agnostic model	✗	✗	✓	✓	✓
Coverage guarantee for test data	✗	✗	✗	✓	✓
Adaption to heteroscedasticity	✓	✓	✓	✗	✓
Computational efficiency	✗	✗	✓	✓	✓

Figure 1: V_{min} prediction flow

for normal chips and inadequate coverage for anomalous ones.

To this end, we propose a distribution-free V_{min} interval prediction framework with a theoretical coverage guarantee. Our approach leverages Conformalized Quantile Regression (CQR) and on-chip monitors to construct prediction intervals. Our primary contributions are outlined as follows:

• We conduct a comprehensive comparison among various V_{min} point predictors for our industrial dataset. We discover that while no golden model outperforms others for all scenarios, the prediction accuracy of linear regression is competitive overall. Moreover, on-chip monitors are capable of predicting future V_{min} degradation.

• We introduce CQR to the context of V_{min} interval estimation, showcasing its better performance in terms of coverage rate and interval length when compared to alternative UQ models.

• Through empirical analysis, we demonstrate that the inclusion of on-chip monitor data yields substantial improvements in the precision of interval predictions.

II. PRELIMINARIES

A. Point Prediction

For the task of V_{min} point estimation, in both the product flow and in-field scenarios, the objective remains consistent: utilizing a set of features to predict a single value. We denote these features as a D dimension vector $\mathbf{x} \in \mathbb{R}^D$, the V_{min} as a real number $y \in \mathbb{R}$, and the point predictor as $g_p(\cdot; \boldsymbol{\theta}) : \mathbb{R}^D \to \mathbb{R}$, parameterized by $\boldsymbol{\theta}$. Given a training dataset of N tested chips $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the predictor is optimized by minimizing the mean of a loss function $\mathcal{L}_p$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_p\big(g_p(\mathbf{X}; \boldsymbol{\theta}), \mathbf{y}\big), \quad (1)$$

where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$ is a matrix of inputs, and $\mathbf{y} = [y_1, \cdots, y_N]^T \in \mathbb{R}^N$ is a vector of true V_{min}.

B. Region Prediction

In manufacturing test processes, engineers often face risks of over-kill or under-kill when relying solely on V_{min} point predictions to identify abnormal products due to process variations. In in-field scenarios, point estimation can be highly unreliable due to the presence of numerous environmental uncertainties.

Consequently, the utilization of prediction intervals becomes essential for effectively detecting outliers and identifying potential failures.

Unlike point estimation, which only generates a single value for an input example, region prediction provides an interval prediction. A region regressor $g_r(\cdot; \boldsymbol{\theta}_{lo}, \boldsymbol{\theta}_{hi}) : \mathbb{R}^D \to \mathbb{R}^2$, consisting of a pair $g_p(\cdot; \boldsymbol{\theta}_{lo}) : \mathbb{R}^D \to \mathbb{R}$ and $g_p(\cdot; \boldsymbol{\theta}_{hi}) : \mathbb{R}^D \to \mathbb{R}$ of the lower and the upper bound function, maps a sample $\mathbf{x}$ to a closed region $C(\mathbf{x})$:

$$C(\mathbf{x}) = \big[g_p(\mathbf{x}; \boldsymbol{\theta}_{lo}), \quad g_p(\mathbf{x}; \boldsymbol{\theta}_{hi})\big]. \quad (2)$$

Given a coverage rate $1 - \alpha$ where $\alpha \in [0, 1]$ and the training dataset $\mathcal{D}$, the prediction intervals of a region regressor g_r should be able to cover at least $1 - \alpha$ labels:

$$\mathbb{P}\big\{y \in C(\mathbf{x}) | (\mathbf{x}, y) \in \mathcal{D}\big\} \geq 1 - \alpha. \quad (3)$$

We introduce two well-known region regression methods satisfying Eq. (3): Gaussian process and quantile regression. Their theoretical traits are summarized in Table I.

1) Gaussian Process (GP): GP is a non-parametric Bayesian method that provides a posterior Gaussian distribution for any testing point [6]. Suppose the posterior mean is $\mu(\mathbf{x}) \in \mathbb{R}$ and the posterior variance is $\sigma^2(\mathbf{x}) \geq 0$ for sample $\mathbf{x}$, we are able to construct an interval $C(\mathbf{x})$ satisfying Eq. (3):

$$C(\mathbf{x}) = \big[\mu(\mathbf{x}) + K_{lo}\sigma(\mathbf{x}), \quad \mu(\mathbf{x}) + K_{hi}\sigma(\mathbf{x})\big], \quad (4)$$

where $K_{lo} = \Phi^{-1}(\alpha/2) < 0$, $K_{hi} = \Phi^{-1}(1 - \alpha/2) > 0$, and Φ is the cumulative distribution function of the standard Gaussian distribution.

2) Quantile Regression (QR): Apart from traditional regression analysis with Mean Square Error (MSE) loss that estimates the conditional mean of V_{min}, QR estimates the conditional quantile [9]. Given a quantile $q \in [0, 1]$, a QR model is trained to minimize the quantile loss [9] $\mathcal{L}_q$ in Eq. (1):

$$\mathcal{L}_q\big(y, \hat{y}\big) := \max \big\{q(y - \hat{y}), (1 - q)(\hat{y} - y)\big\}, \quad (5)$$

where $\hat{y} = g_p(\mathbf{x}; \boldsymbol{\theta})$ is the prediction of quantile V_{min}.

By selecting two different quantiles $q_{lo} = \alpha/2$ and $q_{hi} = 1 - \alpha/2$, we can train two quantile regressors, the interval between which achieves the coverage in Eq. (3).

QR can be easily added to any point regressor where its objective is to minimize the MSE loss by applying the pinball loss instead.

III. METHODOLOGY

A. Overview of V_{min} Prediction

Our V_{min} prediction framework is depicted in Fig. 1, where four stress read points are drawn for illustration. V_{min} at each stress read point will be predicted. The horizontal dash line (min_spec) stands for the product specification of the minimum operating voltage, i.e., device with V_{min} higher than that threshold will violate the specification and likely become a failure.

We utilize low-cost parametric data and on-chip data to predict V_{min} at time zero and subsequent read points during stress simulated in-field life. Note that stress is done at an

elevated voltage such that a much shorter stress duration is equivalent to a much longer in-field life. Specifically, two kinds of V_{min} prediction scenarios are considered: in the production test flow, and in the in-field deployment which is simulated by accelerated stress. In the first case, both production parametric test data and on-chip data are included to build V_{min} predictors. In the second case, however, we make V_{min} degradation prediction based on all accessible features before the V_{min} test timestamp, including production parametric test data at time zero and on-chip data measured at all previous read points during stress. In our industrial dataset, both V_{min} and on-chip data are collected at the same read point, and the total number of read points is relatively small, i.e., less than 10. In this case, time series methods would suffer over-fitting problems. Thus, we treat on-chip data at different read points as different features, and apply CQR to predict V_{min} intervals.

Since CQR is originated from CP, we first briefly summarize how CP works, and then present CQR for V_{min} interval prediction.

B. Conformal Prediction (CP)

Even though the coverage of prediction intervals is guaranteed for the training dataset $\mathcal{D}$ in GP and QR, such characteristic is not held for a testing instance $(\mathbf{x}_{N+1}, y_{N+1})$:

$$\mathbb{P}\{y_{N+1} \in C(\mathbf{x}_{N+1})\} \geq 1 - \alpha. \quad (6)$$

The adoption of the aforementioned two region predictors for new examples is risky without the coverage guarantee.

In semiconductor industry, all chips can be viewed as examples from a hidden distribution: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N+1}$ are sampled i.i.d. from a distribution P_{XY}. CP can help to calibrate any heuristic interval to meet the coverage guarantee in Eq. (6) [10]. CP has two main versions: *full* CP and *split* CP. In regression tasks, full CP needs infinite times of model fitting, rendering it impossible for practical usage. On the contrary, split CP is more computationally efficient with the scarification of splitting the training dataset.

We outline how split CP utilizes a V_{min} point predictor g_p to generate a interval $C(\mathbf{x})$ for $\mathbf{x}$:
• Split the training dataset $\mathcal{D}$ into a new training dataset $\mathcal{D}_{tr}$, and a small calibration dataset $\mathcal{D}_{ca}$ such that $\mathcal{D}_{tr} \cup \mathcal{D}_{ca} = \mathcal{D}$, and $\mathcal{D}_{tr} \cap \mathcal{D}_{ca} = \phi$.
• Fit the point regressor g_p in $\mathcal{D}_{tr}$.
• Compute $\hat{q}$ as the $\lceil (M+1)(1-\alpha) \rceil / M$-th quantile of the conformal score function $s(\mathbf{x}, y)$ of absolute residuals in the calibration set $\mathcal{D}_{ca}$:

$$s(\mathbf{x}, y) = |y - g_p(\mathbf{x}; \boldsymbol{\theta})|, \quad (7)$$

where M is the number of examples in $\mathcal{D}_{ca}$.
• Construct the interval for $\mathbf{x}_{N+1}$, satisfying Eq. (6):

$$C(\mathbf{x}_{N+1}) = \left[g_p(\mathbf{x}_{N+1}; \boldsymbol{\theta}) - \hat{q}, \quad g_p(\mathbf{x}_{N+1}; \boldsymbol{\theta}) + \hat{q} \right]. \quad (8)$$

C. Conformalized Quantile Regression (CQR)

While split CP satisfies the coverage guarantee, the length of predicted intervals is $2\hat{q}$, remaining fixed to different inputs.

Table II: Input feature description

Attribute	Parametric	On-chip (ROD)	On-chip (CPD)
Quantity	1800	168	10
Temperature (°C)	-45, 25, 125	25	80
Read point (hour)	0	0, 24, 48, 168, 504, 1008	

This property may incur overkill for good products and underkill for defective ones. CQR, however, is a variant interval prediction method combining CP and QR together.

We describe the procedures of *split* CQR:
• Split the training dataset $\mathcal{D}$.
• Fit the quantile regressor g_r in $\mathcal{D}_{tr}$.
• Compute $\hat{q}$ as the $\lceil (M+1)(1-\alpha) \rceil / M$-th quantile of the conformal score function $s(\mathbf{x}, y)$ in $\mathcal{D}_{ca}$, where

$$s(\mathbf{x}, y) = \max\{g_p(\mathbf{x}; \boldsymbol{\theta}_{lo}) - y, \quad y - g_p(\mathbf{x}; \boldsymbol{\theta}_{hi})\}. \quad (9)$$

• Construct the interval for $\mathbf{x}_{N+1}$ satisfying Eq. (6):

$$C(\mathbf{x}_{N+1}) = \left[g_p(\mathbf{x}_{N+1}; \boldsymbol{\theta}_{lo}) - \hat{q}, \quad g_p(\mathbf{x}_{N+1}; \boldsymbol{\theta}_{hi}) + \hat{q} \right]. \quad (10)$$

CQR inherits good features of CP and QR, as shown in Table I. It is shown empirically effective in achieving the shortest interval length than CP and QR across 11 datasets while persisting the designed coverage rate [11]. Herein, we adopt it for reliable V_{min} interval prediction.

IV. EXPERIMENTAL RESULTS

A. Industrial Dataset

Our experiments use 156 5nm automotive chips to demonstrate the effectiveness of the proposed V_{min} prediction framework. As shown in Fig. 1, parametric data and on-chip monitor data are considered for V_{min} prediction. We describe how the input features and the output V_{min} are collected.

All 156 chips go through the dynamic Dhrystone stress at elevated voltage in Burn-In (BI) oven for 1008 hours to simulate in-field long-term aging degradation. At specific stress read points, i.e., 0, 24, 48, 168, 504, and 1008 hours, we pause the stress process and 1) test SCAN V_{min}, 2) perform the parametric tests, and 3) collect on-chip monitor data. SCAN V_{min} is tested on Automatic Test Equipment (ATE) tester, at temperatures of -45°C, 25°C, and 125°C. The parametric tests are also performed on ATE tester, including IDDQ, trip IDD, leakage, etc., across all three temperatures. The chip has two types of on-chip monitors: domain sensors which include Ring Oscillator Delay (ROD) sensors and in-situ Critical Path Delay (CPD) sensors. In our experiment, due to hardware and logistic process limitations, ROD is measured on ATE at room temperature (25°C) only while CPD is measured in-situ in BI oven at 80C. We summarize the traits of input features in Table II.

B. Experimental Settings

We illustrate the features used for V_{min} prediction at each read point and the evaluation metrics for point prediction and interval regression. As shown in Fig. 1, for the prediction of V_{min} at time 0, both parametric test data and on-chip monitor
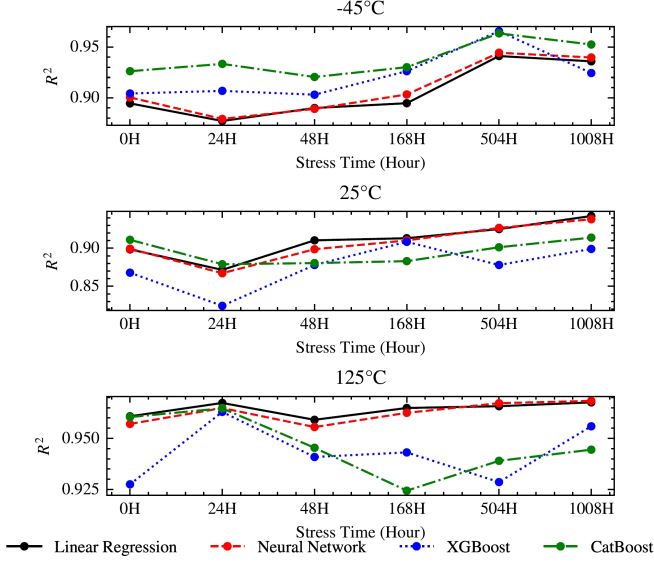
Figure 2: SCAN V_{min} point prediction

data collected at time 0 are utilized to predict V_{min}; For the prediction of V_{min} at the subsequent read points to enable in-field failure prediction, we use on-chip monitor data collected at all previous read points and parametric data collected at time 0, because parametric tests are no longer possible once chips are shipped to customers and deployed in-field.

For V_{min} point prediction, the performance criteria are the coefficient of determination (R^2) and Root Mean Square Error (RMSE); For V_{min} region prediction, the metrics are the average interval length and the coverage of true V_{min} of the testing data.

To reduce the influence of randomization, a 4-fold cross-validation is adopted. We report the average score of each metric across the 4 testing folds. In CQR, 75% training data are used to train predictors while the remaining 25% chips are held for calibration. To ensure a fair comparison, we use the same random seed for all V_{min} interval predictors.

C. Descriptions of V_{min} Point Regressors

ML models with fewer learnable parameters and simpler structures are more favorable for our high-dimensional small data scenario. Moreover, feature selection is an essential dimension reduction technique for some ML models to avoid overfitting problems.

Firstly, we demonstrate model selection for V_{min} point prediction. 5 regressors are considered: Linear Regression (LR), Gaussian Process (GP) [6], XGBoost [12], CatBoost [13], and a 2-layer Neural Network (NN). The detailed configurations of each regressor except LR are provided below:

1) Gaussian Process: GP utilizes a radial basis function kernel, whose parameters are optimized to maximize the likelihood of training data.

2) XGBoost: We utilize the default hyper-parameters in the XGBoost Python package.

3) CatBoost: We utilize the default hyperparameters in the CatBoost Python package except for one hyper-parameter: the number of boosting trees. The default number is 1000, which seems too large for our small dataset including 156 chips, and potentially causes over-fitting. Therefore, we reduce it to 100.

4) Neural Network: We consider a shallow fully-connected multilayer perceptron (MLP) with one hidden layer containing 16 neurons with Rectified Linear Units (ReLU) [14] activation functions. The optimizer is Adam [15] whose learning rate is 0.01, the number of epochs is 3000, and the weight of L_2 penalty is 0.1. These configurations are the same as [5].

Then, we discuss how to select a small set of informative features among thousands of input data. For XGBoost and CatBoost which have an intrinsic feature selection mechanism, all raw data are directly fed to regressors. For the rest of the three methods, we apply Correlation Feature Selection (CFS) [16] with the Pearson correlation to pick 1 to 10 features as input data and report the best testing scores.

D. V_{min} Point Prediction Results

The R^2 of V_{min} point predictions of regression models are depicted in Fig. 2 For SCAN V_{min} tested at time 0, while CatBoost is the best method across all three temperatures, linear regression is also performing well with a small drop of R^2, which is less than 0.03. For all methods except GP, the RMSE for V_{min} point predictions are within $2.5mV$ to $7mV$ ($12mV$ to $22mV$ for GP) for all scenarios, and exhibiting similar comparison as R^2 among different models, i.e., CatBoost performs best for time 0 prediction while linear regression performs reasonably well overall. As linear regression is straightforward to implement by either software or hardware, it is a sufficiently good option for V_{min} time 0 prediction in industrial production tests.

For V_{min} degradation prediction, no regression model is outperforming the rest across all temperatures and stress read points, in terms of R^2 and RMSE. We note that linear regression is still performing reasonably well, and even the best one for predicting SCAN V_{min} at 25°C and 125°C, for both R^2 and RMSE. With its simplicity, implementing a linear regression model with an on-chip hardware accelerator seems to be a viable option for in-field V_{min} degradation prediction.

In addition, an interesting observation is that there is no clear reduction of R^2 in SCAN V_{min} degradation prediction accuracy from 0 to 1008 hours. It demonstrates that our design of on-chip monitors captures informative gate-level features that exhibit a strong correlation with system-level V_{min}.

E. Descriptions of V_{min} Region Regressors

We consider three interval prediction methods: GP, QR, and CQR. QR and CQR are built on 4 point regressors: LR, NN, XGBoost, and CatBoost. The configurations of these models are the same as those in Section IV-C. We set $\alpha = 0.1$ and let predictors generate an interval with 5% to 95% coverage.

F. V_{min} Region Prediction Results

The average length of prediction intervals of SCAN V_{min} and coverage rates are shown in Table III. Both GP and QR

Table III: Average length and coverage of prediction intervals for SCAN V_{min} across 156 chips

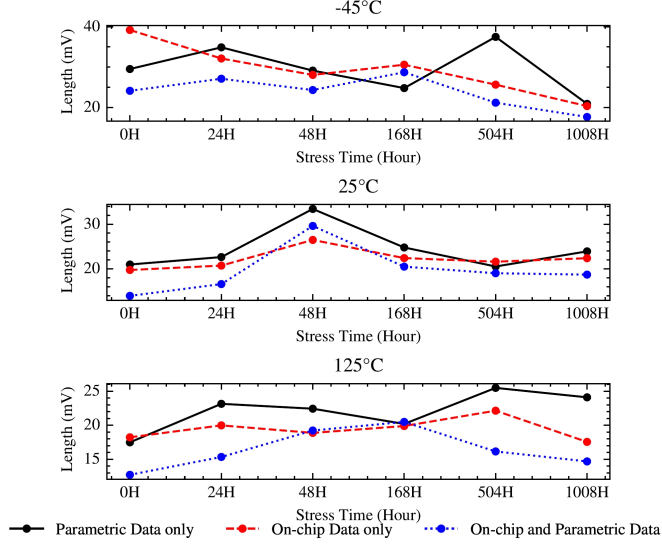| Stress Time (Hour)	Method	-45°C		25°C		125°C	
		Length (mV)	Coverage (%)	Length (mV)	Coverage (%)	Length (mV)	Coverage (%)
0	GP	61.96	85.9	48.56	93.59	51.88	89.1
	QR Linear Regression	51.0	91.03	14.14	83.33	15.98	83.33
	QR Neural Network	30.44	66.84	18.28	53.91	21.33	52.83
	QR XGBoost	50.31	51.28	28.22	89.1	30.96	82.05
	QR CatBoost	2.48	10.26	0.98	14.1	1.37	24.36
	CQR Linear Regression	53.76	92.95	17.37	95.51	19.39	91.03
	CQR Neural Network	114.3	94.81	52.75	93.11	77.54	94.01
	CQR XGBoost	60.84	95.51	31.91	92.95	48.48	98.72
	CQR CatBoost	**24.11**	**91.67**	**13.94**	**92.95**	**12.72**	**91.67**
24	GP	56.76	84.93	48.64	94.87	50.53	87.74
	QR Linear Regression	26.7	85.62	18.3	80.13	13.28	85.16
	QR Neural Network	24.19	68.67	16.33	49.52	19.78	53.68
	QR XGBoost	43.27	39.04	32.64	87.18	30.28	86.45
	QR CatBoost	1.54	3.42	1.38	19.87	1.77	20.65
	CQR Linear Regression	43.1	99.32	20.68	89.74	17.07	95.48
	CQR Neural Network	117.82	97.01	53.66	93.34	84.99	95.45
	CQR XGBoost	65.3	99.32	43.5	92.95	42.41	92.9
	CQR CatBoost	**27.1**	**97.95**	**16.58**	**94.87**	**15.34**	**93.55**
48	GP	56.83	81.13	49.84	89.72	53.84	82.24
	QR Linear Regression	29.77	84.91	20.03	81.31	13.98	82.24
	QR Neural Network	29.66	68.04	44.71	92.05	26.14	50.79
	QR XGBoost	45.43	45.28	35.78	85.98	48.6	84.11
	QR CatBoost	1.64	11.32	1.07	16.82	1.79	19.63
	CQR Linear Regression	36.92	93.4	**29.34**	**94.39**	**20.61**	**93.46**
	CQR Neural Network	100.62	95.59	58.75	95.62	80.64	95.07
	CQR XGBoost	62.81	98.11	49.82	94.39	55.12	95.33
	CQR CatBoost	**24.3**	**95.28**	**29.61**	**96.26**	**19.23**	**89.72**
168	GP	54.45	79.81	50.43	84.91	54.42	85.58
	QR Linear Regression	26.05	81.73	44.0	89.62	12.27	81.73
	QR Neural Network	27.74	72.68	43.56	84.12	26.03	48.32
	QR XGBoost	38.27	75.96	39.89	84.91	49.65	85.58
	QR CatBoost	1.81	19.23	0.71	13.21	1.78	20.19
	CQR Linear Regression	36.28	92.31	51.35	94.34	**17.09**	**89.42**
	CQR Neural Network	82.98	95.33	60.16	95.48	80.99	95.42
	CQR XGBoost	56.65	96.15	48.61	94.34	57.75	92.31
	CQR CatBoost	**28.71**	**93.27**	**20.49**	**91.51**	**20.49**	**92.31**
504	GP	52.61	77.0	52.63	88.46	54.23	79.61
	QR Linear Regression	25.46	83.0	37.71	88.46	26.14	88.35
	QR Neural Network	25.51	70.39	46.33	92.16	48.65	83.49
	QR XGBoost	35.9	78.0	43.14	84.62	47.71	83.5
	QR CatBoost	1.43	12.0	1.54	18.27	2.24	20.39
	CQR Linear Regression	31.2	91.0	45.21	93.27	32.05	94.17
	CQR Neural Network	66.13	93.37	53.44	92.79	72.25	94.76
	CQR XGBoost	46.81	93.0	46.83	87.5	58.74	96.12
	CQR CatBoost	**21.17**	**96.0**	**19.01**	**92.31**	**16.15**	**94.17**
1008	GP	53.18	78.12	52.45	91.84	53.22	82.65
	QR Linear Regression	29.75	88.54	42.63	88.78	32.28	80.61
	QR Neural Network	20.2	50.3	19.89	39.14	31.47	51.9
	QR XGBoost	37.18	79.17	45.19	84.69	46.0	82.65
	QR CatBoost	1.72	17.71	1.64	13.27	1.89	24.49
	CQR Linear Regression	32.3	89.58	47.25	94.9	36.53	91.84
	CQR Neural Network	78.55	98.2	66.8	93.08	65.86	92.25
	CQR XGBoost	44.14	89.58	47.11	91.84	51.44	96.94
	CQR CatBoost	**17.64**	**93.75**	**18.7**	**94.9**	**14.68**	**89.8**

Figure 3: The average interval length of CQR CatBoost for SCAN V_{min} prediction

Table IV: SCAN V_{min} interval prediction via CQR CatBoost averaged across all stress time read points

Feature type	Avg Interval Length (mV)			
	-45°C	25°C	125°C	Average
Parametric	29.44	24.38	22.14	25.32
On-chip	29.32	22.22	19.44	23.66
On-chip and Parametric	**23.84**	**19.72**	**16.43**	**20.00**
On-chip monitor gain	19.02%	19.11%	25.79%	21.01%

underestimate the interval for testing chips, failing to meet the designed coverage rate. CQR, in contrast, successfully calibrates the undercovered interval predictions of QR across all stress read points and temperatures, underscoring the importance of applying conformal prediction for reliable region predictions.

CQR performs differently with different point regression models. The best variant is CQR CatBoost, achieving the shortest intervals with around 90% coverage rate. While LR is competitive for point prediction in Section IV-D, its CQR version predicts larger intervals than CQR CatBoost, especially for SCAN V_{min} at -45°C and 25°C.

G. Benefits of On-chip Monitors

We present evidence supporting the value of on-chip monitor data in the prediction of V_{min} intervals. Fig. 3 illustrates the interval length of CQR CatBoost with three types of feature sets: 1) parametric test data and on-chip monitor data (same to Section IV-F), 2) parametric test data only, and 3) on-chip monitor data only. In addition, Table IV summarizes the average length across all read points of SCAN V_{min} during stress.

Compared to utilizing parametric data only, the inclusion of on-chip monitor data results in a reduction of 21.01% in the average interval length. Intriguingly, a CQR CatBoost model relying solely on on-chip monitor data outperforms the same model using only parametric test data, despite the much larger

number of parametric data (Table II). This implies the on-chip monitor data could contain more information that facilitates V_{min} estimation.

V. CONCLUSION

We propose a distribution-free V_{min} interval estimation framework possessing a statistical coverage guarantee. By harnessing CQR in conjunction with on-chip monitor data, our approach achieves an average interval length of $20mV$ with a 90% coverage rate for true V_{min} values on our industrial dataset. In the future, we will explore how to embed the proposed method 1) in the production test flow to accelerate the V_{min} test and enhance the yield while screening out outliers, and 2) in the in-field systems to secure long-term reliability and safety.

ACKNOWLEDGMENT

REFERENCES

[1] C. He and Y. Yu, "Wafer level stress: Enabling zero defect quality for automotive microcontrollers without package burn-in," in *2020 IEEE International Test Conference (ITC)*, 2020, pp. 1–10.

[2] T.-B. Chan, P. Gupta, A. B. Kahng, and L. Lai, "Ddro: A novel performance monitoring methodology based on design-dependent ring oscillators," in *Thirteenth International Symposium on Quality Electronic Design (ISQED)*, 2012, pp. 633–640.

[3] J. Chen, J. Zeng, L.-C. Wang, J. Rearick, and M. Mateja, "Selecting the most relevant structural fmax for system fmax correlation," in *2010 28th VLSI Test Symposium (VTS)*, 2010, pp. 99–104.

[4] W.-C. Lin, C. Chen, C.-H. Hsieh, J. C.-M. Li, E. J.-W. Fang, and S. S.-Y. Hsueh, "Ml-assisted vminbinning with multiple guard bands for low power consumption," in *2022 IEEE International Test Conference (ITC)*, 2022, pp. 213–218.

[5] Y. Yin, R. Chen, C. He, and P. Li, "Domain-specific machine learning based minimum operating voltage prediction using on-chip monitor data," in *2023 IEEE International Test Conference (ITC)*, 2023, pp. 99–104.

[6] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[7] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-on bayesian neural networks—a tutorial for deep learning users," *IEEE Computational Intelligence Magazine*, vol. 17, no. 2, pp. 29–48, 2022.

[8] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[9] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

[10] G. Shafer and V. Vovk, "A tutorial on conformal prediction." *Journal of Machine Learning Research*, vol. 9, no. 3, 2008.

[11] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," *Advances in neural information processing systems*, vol. 32, 2019.

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," vol. 31, 2018.

[14] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, p. 807–814.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.