Domain-Specific Machine Learning based Minimum Operating Voltage Prediction using On-Chip Monitor Data

Yuxuan Yin
Dept. of ECE
University of California
Santa Barbara, USA
y_yin@ucsb.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*—Determining the minimum operating voltage (V_{min}) of chip designs is critical for low power dissipation and assurance of quality and functional safety during manufacturing tests and in-field monitoring. We demonstrate how on-chip monitor data can be leveraged to provide accurate minimum operating voltage prediction using a domain-specific machine learning approach. Given limited measured chip data, the key challenge in developing a machine learning approach is to provide an accurate prediction while addressing overfitting and selecting a subset of optimal features. To this end, we propose to utilize a novel monotonic lattice neural network architecture that is geared towards accurate prediction by imposing domain-specific monotonic relationships between the input sensor data and V_{min}. Furthermore, we perform an effective feature selection by considering both the correlation between each feature and V_{min} as well as the co-linearity between the features. Experiments demonstrate superior performance in comparison with linear regression and conventional neural networks.**

***Index Terms*—Chip performance prediction, On-chip monitor, Machine learning, Monotonicity**

I. Introduction

The minimum operating voltage (V_{min}) is a critical performance metric for chips, ensuring low power dissipation, quality, and functional safety during manufacturing testing, and in-field monitoring. Excessive energy consumption occurs when the operating voltage exceeds V_{min} [1], and inadequate power supplies may lead to customer returns. Moreover, V_{min} tests using structural test patterns like SCAN and SRAM Built-In Self Test (MBIST), have become more effective and important at detecting subtle defects during manufacturing testing, especially on advanced technology nodes [2].

However, chip performance measurements require complex instrumental setups and long test times, rendering them costly to apply in the production flow and inapplicable to in-field monitoring. Even in production test flow, measuring V_{min} for each test pattern, i.e., running each test pattern at multiple decreasing voltage steps until it fails, will take a huge amount of test time which will become cost-inhibiting. Typically, in current industry practice, the V_{min} search will be run only on limited sample size for a product during the characterization phase to determine a static fixed voltage level V_{min} for each test pattern, and then in the production test, the V_{min} test will be performed at that fixed voltage level V_{min} as a pass / fail test on all chips for that product. Due to the inherent part-to-part process variation, such static fixed V_{min} voltage will inevitably over-reject some chips, i.e., V_{min} test voltage is too low, resulting in unnecessary yield loss; while on some chips it will under-reject, i.e., V_{min} test voltage is too high, resulting potential quality escapes. Therefore, it will be desirable to dynamically adjust the V_{min} test voltage, i.e., V_{min} prediction, for each chip based on on-chip monitors.

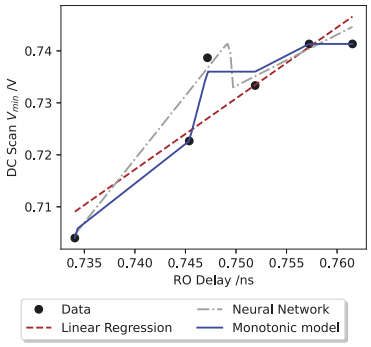


Figure 1: System-level DC scan V_{min} regression

At a high level, there are two types of on-chip monitors: 1) Domain sensors, which typically are a set of Ring Oscillator (RO) based sensors [3]. They are built with the representative devices and logic gates used in each chip such that they will have a good correlation to the chip-level performance and aging degradation when stressed together with the rest of the chip. 2) In-situ critical path margin sensors, which can be implemented also as RO-based sensors or even more light-weight Flip-Flop (FF) based delay margin sensors [4]. They are placed into the selected critical paths on each chip to provide the measurement of the critical path delay margin and degradation over stress directly, i.e., in-situ. In [5], a good

2023 IEEE International Test Conference (ITC) | 979-8-3503-4325-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITC51656.2023.00024

2378-2250/23/$31.00 ©2023 IEEE
DOI 10.1109/ITC51656.2023.00024

Short Paper

correlation of the on-chip monitor data and chip-level V_{min} has been demonstrated and linear regression is used to predict V_{min} from on-chip monitor data with limited accuracy.

To this end, ML approaches using on-chip monitor data are more promising for V_{min} prediction. However, ML methods encounter challenges of limited expensive measured chips and high-dimensional input features: a typical artificial neural network could easily overfit. Thus, previous approaches seldom adopted neural networks as their predictors but employed less capable alternatives instead, including linear regression [6], [7], Gaussian process [8], and support vector machine [9]. However, none of them have domain knowledge embedded. On-chip data features, such as measurements of RO delay and path-based RO delay [4], are correlated with the chip performance in a specific way, as a degraded chip unit never leads to better performance. For instance, an increased RO delay cannot yield a decreased V_{min}. This physical constraint can be mathematically formulated as *monotonicity* between the input features and the targeted V_{min}, and its rigorous definition is presented in Section III-C. Fig. 1 shows the DC scan V_{min} prediction by linear regression, a conventional neural network, and a monotonic neural network model based on the proposed approach using one RO delay as an input feature. As the RO delay increases, V_{min} should be non-decreasing, but the conventional neural network violates this rule. On the other hand, linear regression has inferior performance because of its limited predictive capacity.

In this paper, we introduce a novel domain-specific framework for predicting V_{min} using a lattice network [10] that embeds monotonic constraints to address overfitting issues. Furthermore, to eliminate redundancies in highly correlated on-chip data while retaining useful information, we present a feature selection approach that selects features with high monotonic correlations to V_{min} and low linear dependencies among themselves. We demonstrate the superior performance of our method on 5nm automotive microprocessor chips. In addition to on-chip RO monitors capturing chip-level properties, multiple in-situ critical path delay monitors are implemented to provide local aging degradation of chips. The results demonstrate the effectiveness of our approach, and the main contributions of our work are summarized as follows:

- We propose a domain-knowledge embedded monotonic lattice network based V_{min} predictor, which avoids over-fitting due to limited training data from tested chips.
- We present a feature selection method that is specifically designed for our monotonic lattice network approach, selecting effective features based on their correlations with V_{min} and linear independence among themselves.
- We show through empirical evidence that critical path delays are effective features for V_{min} prediction.

II. Background

Estimating chip performance, including V_{min} and F_{max} (the maximum operating frequency), has been a topic of intense research. Lin proposed a V_{min} prediction flow that includes multiple guard bands to account for both underprediction and overprediction [1]. The authors employed linear regression to predict the system V_{min}, logistic regression to classify chips into different bins based on prediction error, and guard bands to each bin. Kuo proposed a robust framework to predict V_{min}, taking into account the variations from lot to lot [6]. The authors trained an initial linear regression model for the prediction of V_{min} and updated it on chips from new lots by accumulative learning when the prediction accuracy fell below a certain threshold. Although effective, the aforementioned approaches had limited capacity compared to neural networks, and none of them took into account the domain-specific monotonicity constraints between features and chip performance.

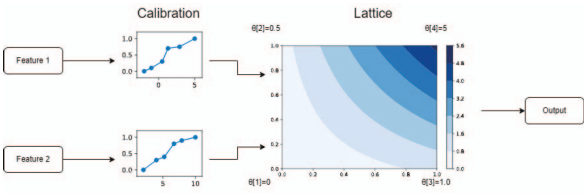


Figure 2: An lattice network with 2 input features.

Several methods have been proposed to embed monotonicity constraints into neural networks [11]–[13]. However, the model in [11] was only able to serve a single input, and no strict monotonicity constraint was guaranteed in [12], thus limiting their usages. In contrast, the monotonic lattice network [13] is a better choice: it strictly follows monotonic constraints and is able to deal with multiple input features. Hence, we adopt it as the V_{min} predictor in our framework.

III. Proposed Method

A. Problem Formulation

We suppose that the features of each chip collected by on-chip monitors are the same and their number is D. This assumption is valid when all chips have the same design. For every chip, its features are represented as a vector $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^D$, and its corresponding V_{min} value is represented as a scalar $\mathrm{y} \in \mathcal{Y} \subset \mathbb{R}$. Here, $\mathcal{X}$ and $\mathcal{Y}$ indicate the input and output spaces, respectively.

We aim to build an ML-based V_{min} predictor $g(\cdot;\boldsymbol{\theta}) : \mathcal{X} \to \mathcal{Y}$, parameterized by $\boldsymbol{\theta}$. To achieve this objective, a training dataset of N tested chips $\mathcal{D} = \{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^{N}$ is used, where $\mathbf{x}_i$ represents the features of the i-th chip, and y_i represents its corresponding V_{min} value. The optimal parameters $\boldsymbol{\theta}^*$ are obtained by minimizing the mean of a loss function $\mathcal{L}$ on $\mathcal{D}$

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathcal{L}\big(g(\mathbf{x}_i;\boldsymbol{\theta}), \mathrm{y}_i\big). \tag{1}$$

B. Overview of Proposed Framework

In order to leverage the power of neural networks while avoiding overfitting on a limited number of tested chips, we propose a novel framework for V_{min} prediction that incorporates domain knowledge that the chip performance is monotonically correlated to some features. Specifically, we adopt a

monotonic lattice network as a V_{min} predictor to fulfill these constraints. Furthermore, we trick the conventional Correlation Feature selection [14] to eliminate redundant inputs while maintaining features monotonically correlating to V_{min}.

C. Monotonic Lattice Network

Domain knowledge states the monotonicity between specific features and V_{min}. The rigorous definition of monotonicity is presented below.

***Definition** 1:* A function $f : \mathcal{X} \rightarrow \mathcal{Y}$ is said to be *monotonically increasing* with respect to a feature set S if $f(\mathbf{x}_i) \geq f(\mathbf{x}_j)$ holds for all input points $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ such that

$$\mathbf{x}_i[k] \geq \mathbf{x}_j[k], \ \forall \ k \in S,$$
$$\mathbf{x}_i[k] = \mathbf{x}_j[k], \ \forall \ k \notin S,$$

where $\mathbf{x}_i[k]$ and $\mathbf{x}_j[k]$ represent the kth feature of point $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively.

We provide a concise overview of the monotonic lattice network, built upon lattice regression [15]. Lattice regression applies multilinear interpolation to fit a function. For a D dimensional input space $[0,1]^D$, a D dimensional unit lattice is used to regress via its 2^D vertices, whose values are denoted by a vector $\boldsymbol{\theta} \in \mathbb{R}^{2^D}$. Given input $\mathbf{x} \in [0,1]^D$, the output of the lattice is $\boldsymbol{\theta}^T \psi(\mathbf{x})$, where $\psi(\cdot)$ is a kernel function. $\psi : [0,1]^D \rightarrow [0,1]^{2^D}$ is a non-linear function whose jth entry $\psi(\mathbf{x})[j]$ satisfies

$$\psi(\mathbf{x})[j] = \prod_{i=1}^{D} \mathbf{x}[i]^{\boldsymbol{v}_j[i]} (1 - \mathbf{x}[i])^{1-\boldsymbol{v}_j[i]}, \tag{2}$$

where $\boldsymbol{v}_j[i] \in \{0,1\}$ is the ith entry of the jth vertex whose coordinate is represented by the vector $\boldsymbol{v}_j$.

Gupta proved what constraints are sufficient and necessary for lattice regression to become monotonic with respect to some features [13]:

***Proposition** 1:* Lattice regression in Eq. (2) is *monotonically increasing* with respect to the feature set S if and only if for all verteices $\boldsymbol{v}_i$, $\boldsymbol{v}_j$, $i, j \in \{1, 2, \cdots, 2^D\}$, feature $\mathbf{x}[k] \in S$, $\boldsymbol{\theta}[i] \geq \boldsymbol{\theta}[j]$ holds such that $\boldsymbol{v}_i[k] = 1$, $\boldsymbol{v}_j[k] = 0$, and $\boldsymbol{v}_i[k'] = \boldsymbol{v}_j[k']$ for all $k' \in \{1, 2, \cdots, D\} \setminus k$.

The aforementioned lattice regression requires input within the unit cube $[0,1]^D$, which limited its usage. Thus, a monotonic piecewise calibrated transformation $c_i : \mathbb{R} \rightarrow [0,1]$ is applied to each feature $\mathbf{x}[i]$, $i \in \{1, 2, \cdots, D\}$ [13]. The monotonicity is preserved after the transformation. An example of a lattice network with two monotonically increasing features is depicted in Fig. 2.

Given the training dataset $\mathcal{D}$, the best parameters $\boldsymbol{\theta}^*$ can be derived from Eq. (1) by minimizing a Mean Square Error (MSE) loss.

D. Feature Selection

Feature selection techniques are commonly used when input features exhibit high redundancy. By selecting a representative subset of features, we can improve prediction accuracy and mitigate the problem of overfitting. we first review a well-known feature selection method named Correlation Feature Selection (CFS) [14], then we propose an adjusted variant to better facilitate monotonic predictors.

CFS picks good representative features regarding the output with small redundancy. Given the ith feature vector $\mathbf{x}[i] = [\mathbf{x}_1[i], \mathbf{x}_2[i], \cdots, \mathbf{x}_N[i]]^T$, the target vector $\mathbf{y} = [\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_N]^T$ in dataset $\mathcal{D}$, and a correlation metric s, a feature subset S of K features are selected by CFS via solving an optimization problem

$$S = \underset{|S|=K}{\arg\max} \frac{\sum_{j \in S} s(\mathbf{y}, \mathbf{x}[j])}{\sqrt{K + \sum_{i,j \in S} 2s(\mathbf{x}[i], \mathbf{x}[j])}}. \tag{3}$$

One is able to filter a set of linear and non-linear features by applying the Pearson Correlation Coefficient (PCC) [16] and Mutual Information (MI) [17], respectively.

Even though theoretically all critical path delays and RO delays are monotonic to V_{min}, the empirical data do not strictly obey this rule because of manufacturing variations and measurement errors. Thus, selected features are supposed to have better monotonicity to V_{min} However, features selected by vanilla CFS with PCC or MI are not guaranteed to have good monotonic correlations to V_{min}.

To this end, we propose Monotonic CFS (M-CFS), which adopts Spearman Correlation Coefficient (SCC) [18], a monotonic metric, for feature-V_{min} correlations.

SCC $s_S(\mathbf{y}, \mathbf{x}[i])$ computes the PCC score s_P between the rank of variables $\mathbf{x}[i]$ and $\mathbf{y}$:

$$s_S(\mathbf{y}, \mathbf{x}[i]) = s_P(r(\mathbf{y}), r(\mathbf{x}[i])), \tag{4}$$

where the rank function r projects each sample to its ascending order: $r(\mathrm{y}_j) = |\{\mathrm{y}_i : \mathrm{y}_i \leq \mathrm{y}_j, \ i \in \{1, \cdots, N\} \setminus j\}|$.

Finally, we present M-CFS. Unlike simply choosing SCC as the correlation metric in CFS, we utilize SCC for feature-V_{min} correlations and PCC for feature-feature correlations:

$$S = \underset{|S|=K}{\arg\max} \frac{\sum_{j \in S} s_S(\mathbf{y}, \mathbf{x}[j])}{\sqrt{K + \sum_{i,j \in S} 2s_P(\mathbf{x}[i], \mathbf{x}[j])}}. \tag{5}$$

The inspiration is that we want to maximize the monotonicity between features and the target, but minimizing the monotonicity between the features is too strict. Therefore, we relax the objective of minimizing the linear dependency among features.

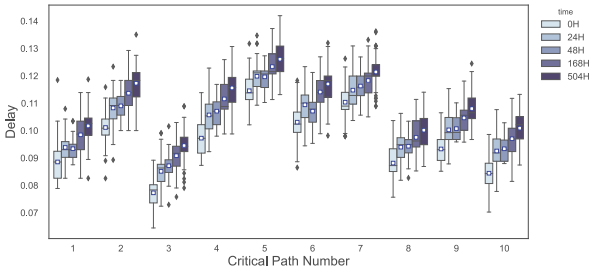
IV. Experimental Results

A. Dataset Description

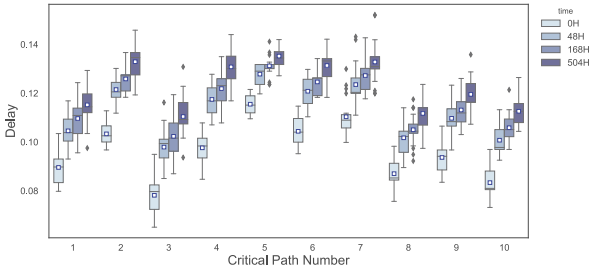
We make use of 96 advanced 5nm automotive test chips to demonstrate how our ML-based method can leverage on-chip monitor data to predict V_{min}. This 5nm automotive test chip contains two types of on-chip monitors: 168 RO-based

Table I: Industrial Chip Dataset

Stress	# Sample	# CPD	# ROD	# V_{min}	# Timestamp
Dhrystone	72	10	168	9	5
HTOL	24	10	168	9	4



(a) Dhrystone stress



(b) HTOL stress

Figure 3: Degradation of 10 critical path delays through Dhrystone stress and HTOL stress.

Delay (ROD) domain sensors, and 10 FF-based in-situ Critical Path Delay (CPD) sensors. After passing structural tests and functional tests on Automotive Test Equipment (ATE), 72 chips have gone through Dhrystone stress while the other 24 chips encounter High-Temperature Operating Life (HTOL) stress. Both Dhrystone and HTOL stress are performed at elevated voltage and temperature on Burn-In (BI) boards inside the BI oven to accelerate the aging of devices.

At specific stress timestamps, we pause the stress and then measure chips. For Dhrystone stress, we test V_{min} at 0 hours, 24 hours, 48 hours, 168 hours, and 504 hours read points, while 24 hours read point was skipped in the test of HTOL stress. V_{min} and RO-based domain sensor data are measured at 25°C on ATE, while the CPD sensor data are only collected in-situ within BI at around 80°C. Table I summarizes our industrial dataset.

B. Benefits of Critical Path Delay for V_{min} Degradation

In this section, we use experimental results to demonstrate the benefits of on-chip in-situ CPD monitors. Firstly, we illustrate that CPD monitors are able to capture the aging degradation of chips. In Fig. 3, delay shifts of 10 critical paths through Dhrystone and HTOL stress are depicted as box plots. Each box plot is corresponding to the scaled delay distribution for each selected critical path from 72 chips that received Dhrystone stress and 24 chips that received HTOL stress, in which the bar inside the box represents the *median*. The box shows the quartiles of the dataset while

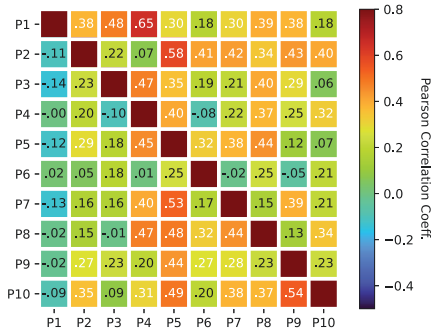


Figure 4: Pearson correlation heatmap of critical path delays. The lower and upper triangular matrix represents the correlations under Dhrystone stress and HTOL stress, respectively.

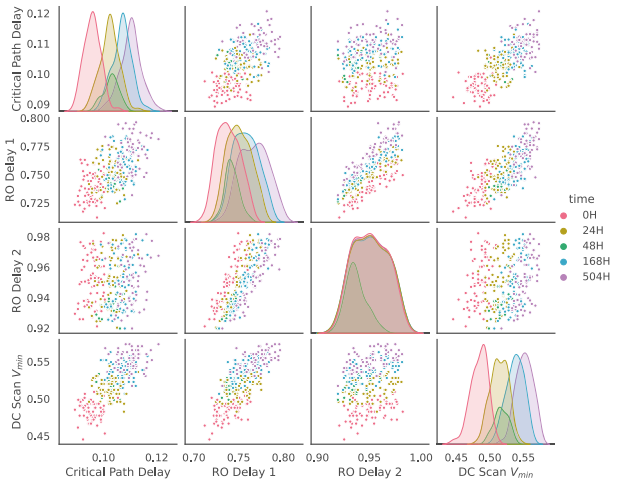


Figure 5: Visualization of the pair-wise correlation among critical path delay, RO delay, and DC scan V_{min} of 72 chips under Dhrystone stress.

the whiskers extend to show the rest of the distribution. The lower whisker extends to 1.5 times interquartile outside Q1 (25th percentile), and the upper whisker extends to 1.5 times interquartile outside Q3 (75th percentile). The square dots are corresponding to the means of the scaled delay distribution for each critical path. The mean and median delay of each path is monotonically increasing with respect to the stress, and it satisfies the domain constraint. Moreover, the rate of delay increase gets slower when the stress time gets longer, indicating a chip-level aging property. It is also noteworthy that the CPD shifts over stress are different among different critical paths which are expected as those different critical paths consist of different types of devices and logic gates. In Fig. 4, we plot the Pearson correlations among CPDs through Dhrystone and HTOL stress, where the linear dependency is low for most critical paths. Hence, CPD monitors capture the aging characteristics of each selected critical path with

Table II: V_{min} Degradation Prediction RMSE of LR

Test Pattern	1 ROD	2 RODs	3 RODs	2 RODs + mean CPD
DC Scan	14.3mV	14.2mV	14.5mV	**12.0mV**
AC Scan	9.97mV	9.81mV	10.0mV	**9.52mV**
MBIST	13.9mV	13.3mV	13.3mV	**13.2mV**

Table III: The Percentage of CFS-selected Features Containing at least 1 CPD for V_{min} Degradation Prediction

Test Pattern	1 Feature	2 Features	3 Features
DC Scan	58.3%	91.7%	100%
AC Scan	48.6%	75.0%	91.7%
MBIST	0%	0%	33.3%

robustness.

Secondly, we show that CPDs provide additional information on V_{min} degradation to RODs. In Fig. 5, we depict the relationship between 1 CPD, 2 RODs, and the DC scan V_{min} tested at 125°C, where the CPD is the *average* scaled delay among the 10 critical paths and ROD 1 and ROD 2 are selected by CFS with the PCC score. As shown in Fig. 5, CPD has an apparent correlation with DC Scan V_{min}, while there is less correlation between CPD and ROD 1 and no correlation between CPD and ROD 2.

Thirdly, we demonstrate the effectiveness of CPD in predicting V_{min} degradation from the 72 chips going through Dhrystone stress. We use on-chip monitor data collected at one stress timestamp to predict V_{min} measured at the *same* timestamp for all test patterns. In order to ensure a fair comparison, we compare the prediction accuracy of linear regression using different on-chip monitor data sets, including RODs and CPD. RODs are determined by CFS in Eq. (3) using PCC. The CPD here is the *mean* delay among 10 critical paths. In order to avoid random fluctuation, we perform 4-Fold cross-validation, train the linear predictors on 54 random chips (75%), and test them on the remaining 18 chips (25%).

We report the Root Mean Square Error (RMSE) of V_{min} prediction averaged across temperatures and timestamps in Table II. If we only use RO-based domain monitor data, 2 selected RODs can be adequate to train a linear V_{min} predictor, and adding one more ROD will even induce a worse performance. However, Our results show that adding CPD data improves the prediction accuracy for DC Scan V_{min}, AC Scan V_{min}, and MBIST V_{min} by 15.7%, 2.98%, and 0.76%, respectively. This indicates that critical path delay is a useful feature in addition to RO delay for V_{min} degradation prediction, especially for DC Scan V_{min}, which is expected as the CPD sensors add visibility of performance margin at each local critical path level on top of the chip-level process feedback from the RO-based domain sensors.

Furthermore, we present evidence that CPDs can be chosen as useful features for V_{min} degradation prediction by CFS. We apply CFS with Pearson correlation s_P as the score function in Eq. (3) to determine the first, the second, and the third feature from CPDs and RODs. We define a successful pick if a selected feature set contains at least one CPD, and display the successful pick percentage for each V_{min} test pattern in Table III. Our results show that CPD has a high chance of being selected as the second or the third feature for DC Scan V_{min} and AC Scan V_{min}. However, CPDs are less compatible for MBIST V_{min}, which is consistent with the result shown in Table II that 2 linear models using 3 RODs or 2 RODs +1 CPDs have similar performance. To this end, we empirically demonstrate that CFS can determine beneficial CPDs for V_{min} degradation prediction.

Table IV: Configurations of V_{min} Predictor

Model	Training epochs	Batch size	Optimizer	Learning rate
MLP	3000	64	Adam	0.001
MLN	8000	64	Adam	0.01

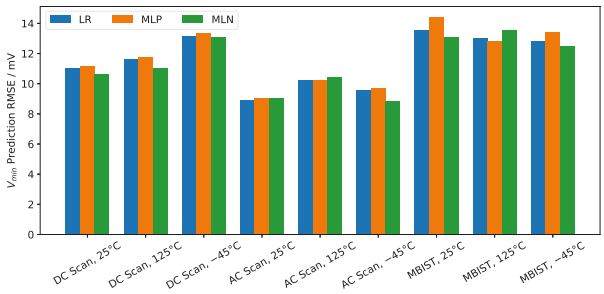


Figure 6: V_{min} prediction RMSE.

C. Effectiveness of the Proposed V_{min} Prediction Framework

We compare our domain-specific ML approach to two baselines: a linear regression model (LR) and a conventional neural network, the latter of which is a multilayer perceptron (MLP). In order to mitigate the overfitting phenomenon, we instantiate a shallow MLP that has one hidden layer including 16 neurons, whose activation functions are Rectified Linear Units (ReLU). Moreover, a L_2 penalty of model parameters with a weight of 0.1 is added to the optimization objective in Eq. (1). Detailed hyper-parameters of the MLP and the monotonic lattice network (MLN) are summarized in Table IV.

We demonstrate that embedding domain-specific monotonicity constraints are beneficial for overfitted neural networks when the dataset is small via two experiments: predicting V_{min} of unstressed chips tested at 0 hours, and modeling V_{min} degradation over stress.

Table V: Prediction R^2 for DC Scan V_{min} Tested at 0 hours Using 3 On-chip Monitor Data Selected by CFS

Testing Temperature	LR	MLP	MLN	MLN with M-CFS
25°C	0.688	0.645	0.711	**0.716**
125°C	0.688	0.686	0.695	**0.699**
-45°C	0.382	0.336	0.393	**0.397**
Mean	0.586	0.561	0.599	**0.604**

Table VI: V_{min} Degradation Prediction RMSE

Test Pattern	LR	MLP	MLN with M-CFS	DK Gain
DC Scan	11.9mV	12.1mV	**11.6mV**	4.13%
AC Scan	9.56mV	9.70mV	**9.46mV**	2.47%
MBIST	13.1mV	13.6mV	**13.0mV**	4.42%

Firstly, we present empirical results of predicting DC Scan V_{min} in 96 chips using LR, MLP, and MLN models. To select input features, we apply CFS with Pearson correlation s_P as the score function in Eq. (3) to choose three on-chip monitor data for each V_{min} from all on-chip monitor data including 168 RODs and 10 CPDs. In addition, we employ M-CFS in Eq. (5) to select three different on-chip monitor data for the MLN. Each predictor is evaluated via a 4-Fold cross-validation.

As shown in Table V, the MLP has lower accuracy for DC scan V_{min} than the LR due to overfitting. However, our proposed MLP predictor that leverages monotonicity constraints provides superior performance. Furthermore, the M-CFS method further improves the performance of the MLP. The results show that our domain-specific ML framework helps improve the accuracy of predicting V_{min} even with a small number of chips tested.

In addition, we demonstrate that our approach is effective in modeling the V_{min} degradation of 72 chips under Dhrystone stress. For each V_{min} test pattern, we perform CFS to select 3 features from the on-chip monitor data. Similar to previous experiments, we use 4-Fold cross-validation to validate the results.

The V_{min} degradation prediction RMSE averaged on timestamps is depicted in Fig. 6. In general, the domain-specific MLN shows higher prediction accuracy than MLP, particularly for DC Scan V_{min} and MBIST V_{min}. Furthermore, we report the average V_{min} prediction RMSE of each test pattern, i.e., average from three temperatures, in Table VI, in which the last column is the improvement percentage of embedding Domain Knowledge (DK gain) using our proposed MLN with M-CFS method comparing to MLP. The results suggest that embedding monotonicity constraints in neural networks is a promising technique for modeling V_{min} degradation even when using only a limited number of measured chips.

V. Conclusion

We introduce a novel framework for predicting V_{min} and its aging degradation using machine learning based on on-chip monitor data. By incorporating domain-specific monotonicity constraints, the accuracy and robustness of the predictions can be significantly improved when measured chips are limited (due to long duration of the stress). Experimental results on advanced 5nm automotive test chips with both on-chip domain sensors and critical path delay monitors have demonstrated the effectiveness of the proposed MLN method. In the future, we will further validate the effectiveness of our approach once a larger volume of industrial stress data is available. Moreover, we will develop an advanced V_{min} degradation predictor based on our proposed MLN method, which is able to predict future V_{min} shifts using past and present on-chip monitor data. It will help enable the prediction of future aging-induced failure, indicated by abnormal V_{min} shift, with a shorter duration of stress needed.

Acknowledgment

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1956313, and partly supported by a NXP Long Term University (LTU) grant.

References

[1] 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.

[2] 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.

[3] J. Keane, W. Zhang, and C. H. Kim, "An array-based odometer system for statistically significant circuit aging characterization," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 10, pp. 2374–2385, 2011.

[4] 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.

[5] J. Todd, M. William, and L. Sharon, "A vmin temperature shift screen for test cost reduction," in *2017 IEEE International Test Conference (ITC)*, 2017.

[6] Y.-T. Kuo, W.-C. Lin, C. Chen, C.-H. Hsieh, J. C.-M. Li, E. Jia-Wei Fang, and S. S.-Y. Hsueh, "Minimum operating voltage prediction in production test using accumulative learning," in *2021 IEEE International Test Conference (ITC)*, 2021, pp. 47–52.

[7] 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.

[8] J. Chen, J. Zeng, L.-C. Wang, M. Mateja, and J. Rearick, "Predicting multi-core system fmax by data-learning methodology," in *Proceedings of 2010 International Symposium on VLSI Design, Automation and Test*, 2010, pp. 220–223.

[9] Q. Shi, X. Wang, L. Winemberg, and M. M. Tehranipoor, "On-chip sensor selection for effective speed-binning," *Analog Integrated Circuits and Signal Processing*, vol. 88, pp. 369–382, 2016.

[10] S. You, D. Ding, K. Canini, J. Pfeifer, and M. Gupta, "Deep Lattice Networks and Partial Monotonic Functions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville, "Neural autoregressive flows," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2078–2087. [Online]. Available: https://proceedings.mlr.press/v80/huang18d.html

[12] A. Gupta, N. Shukla, L. Marla, A. Kolbeinsson, and K. Yellepeddi, "How to incorporate monotonicity in deep networks while preserving flexibility?" *arXiv preprint arXiv:1909.10662*, 2019.

[13] M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, and A. Van Esbroeck, "Monotonic calibrated interpolated look-up tables," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3790–3836, 2016.

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

[15] E. Garcia and M. Gupta, "Lattice regression," *Advances in Neural Information Processing Systems*, vol. 22, 2009.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[17] R. B. Ash, *Information theory*. Courier Corporation, 2012.

[18] J. Hauke and T. Kossowski, "Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data," *Quaestiones geographicae*, vol. 30, no. 2, pp. 87–93, 2011.