Proceedings of the 5th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'19)
Lisbon, Portugal – August, 2019
Paper No. EEE 112
DOI: 10.11159/eee19.112

Minimum Voltage Prediction Model for Application Processor Based on Deep Neural Network in Manufacturing Process

Seunggee Lee[1,3], Jee-Hyong Lee[2], Jungin Han[3]
[1]Dept. of Semiconductor Display Engineering, Sungkyunkwan University
Suwon, Korea
lsk990@skku.edu
[2]Dept. of Software Engineering, Sungkyunkwan University
Suwon, Korea
john@skku.edu
[3]Platform Development Team, System LSI Division, Samsung Electronics Co., Ltd.
Hwaseong, Korea
jihan@samsung.com

Abstract **-** As the functions of mobile devices are diversified, the power consumption of System on Chip (SoC) is increasing rapidly. Among the various types of SoC, Application Processor (AP) controls all functions of mobile device and is the most power-consuming SoC. In the AP manufacturing process, evaluating the characteristics of semiconductor in order to apply the optimal voltage that allows AP to have the lowest power consumption has a great influence on the AP quality. However, it is practically impossible to measure the lowest voltage for individual APs due to considerations such as yield and evaluation time during the manufacturing process. Therefore, voltage binning technique that applies the same voltage by dividing APs into groups according to semiconductor characteristics has traditionally been used, but this can cause a large difference from the lowest voltage that individual APs can actually reach. Therefore, a new voltage prediction model is needed to reduce the over-supplied voltage and accurately estimate the lowest voltage for individual APs. In this paper, we propose a minimum voltage prediction model based on Deep Neural Network (DNN) to evaluate the lowest voltage of APs. We use the leakage current, the ring oscillator and the operation temperature, which are characteristic values of each AP, as inputs to create our prediction model that considers the relationship between each element and the minimum voltage. By using our model, we can find and apply the optimal voltage to each AP so that it operates at lower voltage than the traditional method. Ultimately, we have demonstrated that using the new model can improve the applied voltage by up to 6.76%.

Keywords**:** Deep neural network (DNN), Semiconductor manufacturing, Semiconductor characteristic, Voltage binning, Ring oscillator, Leakage current, System on chip, Application processor.

1. Introduction

As the use of mobile devices such as smartphones increases and performs various functions, usage of APs, which act as the brain of the devices, is going up significantly. In recent years, the role of the AP has become more important as the Neural Processing Unit (NPU) for Artificial Intelligence is added. The amount of AP power consumption is also increasing due to various functions [1]. Therefore, operating voltage optimization for the minimum power consumption is a constant challenge to improve the quality of APs. This is also very important for customer satisfaction and profitability. For the reasons above, low power is a key indicator of AP sales and quality.

There have been various studies to optimize the applied voltage for APs to achieve low power. Most studies have involved Voltage Binning (VB) to optimize power and yield by reducing the variation in SoC characteristics. This method divides APs into Voltage Group (VG) according to semiconductor characteristics and assigns the voltage to APs in each group. Lichtensteiger et al. proposed Selective Voltage Binning (SVB) scheme that maximized yield by reducing the maximum voltage of the Fast Characteristic CMOS SoC. This method mainly reduced the average power consumption by decreasing the dynamic power of Fast CMOS SoC which has a high leakage power by a certain ratio [2]. Zolotov et al. proposed statistical technique of yield computation and this study formulates the problem of computing optimal supply voltages for a proposed binning scheme [3]. Shen et al. described new VB formulation to predict the maximum number of

bins required [4]. However, those studies still use VB to solve the problem, it may differ from the actual lowest voltage in a method in which the same voltage is applied per each group, and average power consumption may not be optimized enough.

In the AP manufacturing process, VB technique has also been used traditionally for voltage optimization and it has used a small number of characteristic parameters to define the applied voltage. Therefore, process has the same limitation of VB scheme as mentioned above. For these reasons, using traditional method does not accurately define the minimum voltage.

In order to solve the above problems, we propose a minimum voltage prediction model based on DNN. The DNN algorithm is adopted because it can analyze and learn the data considering the complex nonlinearity relationship between the characteristic parameter and minimum voltage of each model (V_{min}). We trained our model using the actual measured minimum voltage ($V_{min\_measured}$) and characteristic parameters of the assembled AP. In order to obtain the best results in consideration of various situations, we trained the model by power domain and level, respectively.

Furthermore, by precisely predicting the V_{min} with our proposed model, the test time burden on the lowest voltage search is reduced, which can give good results in reducing the manufacturing cost. Finally, AP power consumption can be reduced by maximum 6.76% per power domain frequency(speed) level, average 3.97% for all domains, and this process can contribute to improve AP quality.

2. Minimum voltage prediction model

In this session, we describe the components of the proposed model. Our model predicts the V_{min} by finding the correlation between $V_{min\_measured}$, which is measured in the initial evaluation of mass production process, and the characteristic elements that represent each AP.

2.1. Input data

Leakage Current (LC), Ring Oscillator (RO) and operation temperature (T_{oper}) are used as input of the prediction model.

RO represents the RO logic delay time measured in the wafer test, and it is stored in the Fuse-box inside the AP. The RO value can be obtained for each power domain. Each value serves as an index for its power domain performance [5]. In our proposed model, we used 10 RO inputs that best represent the performance characteristics of the power domain. To compensate for the difference in the RO value by position, we used 5 RO deviations per position as inputs.

LC is the current flowing in CMOS transistors in static status, and the value is measured differently according to the performance variation of the APs. LC value is high in fast process AP and low in slow process AP. 1 LC value related to the power domain was used as the basic input, and 2 LC values related to the AP internal interface domain were used as additional inputs to improve the accuracy.

T_{oper} has a nearly linear relationship with leakage power, so leakage rises with increasing temperature [6]. Therefore, we used T_{oper} as an input in case the AP temperature difference may affect V_{min}.

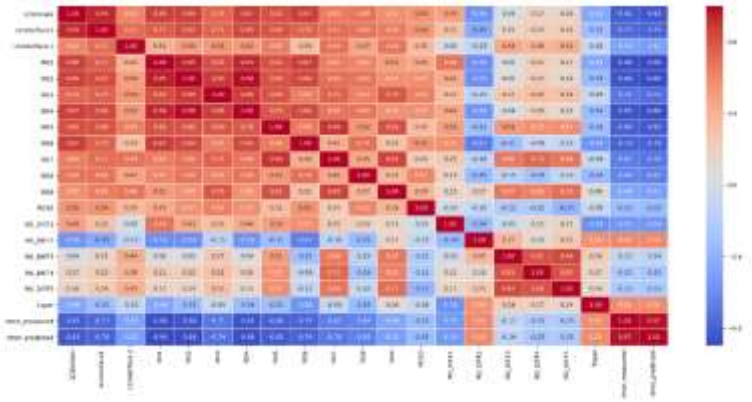


Fig. 1: Correlation matrix of characteristic element and V_{min} in Power Domain *A*, Frequency Level 0.
Deep blue and red color represent high association between each item.

Fig. 1 shows the correlation matrix of each variable. By analyzing correlation between all characteristic elements, we selected input variables to help predict the correct V_{min} and to optimize the training cost.

2.2. Deep Neural Network (DNN)

We adopted the DNN algorithm as a prediction model to solve the problem considering the complex nonlinearity relationship between V_{min} and input value. DNN is one of the machine learning algorithms and is a hierarchical network model using neurons. The proposed DNN model consists of an input layer with 19 nodes, 3 hidden layers, and a voltage prediction output layer. Each layer is fully connected. To avoid overfitting and to take advantage of learning time, DNN can be constructed with only three hidden-layers and each hidden layer needs 2/3 of input layer nodes [7]. Therefore, our DNN structure using 19 inputs is shown in Fig. 2.

There are various kinds of activation functions of hidden layer nodes. A sigmoid, the simple activation function, using only 0 and 1 as the node output can cause a vanishing gradient problem. Therefore, Relu, Elu, and Selu have been proposed as alternatives to avoid the problem and to ensure learning stability. We used Selu in our prediction model. The reason for using Selu among several activation functions was that it had higher stability and lower validation loss than other activation functions with our dataset [8].

DNN is trained to minimize loss function values using the Gradient Descent algorithm. In our predictive model, we used Mean Square Error (MSE), commonly used in regression problems. We chose Adam optimizer, which is widely used for optimizing the process of finding the parameter of the loss function [9].

In addition, we set different epochs for each power domain and frequency level in order to gain the flexibility of the experiment and to increase the accuracy of each model. For this reason, early stopping was used to prevent overfitting when training each frequency level model [10].

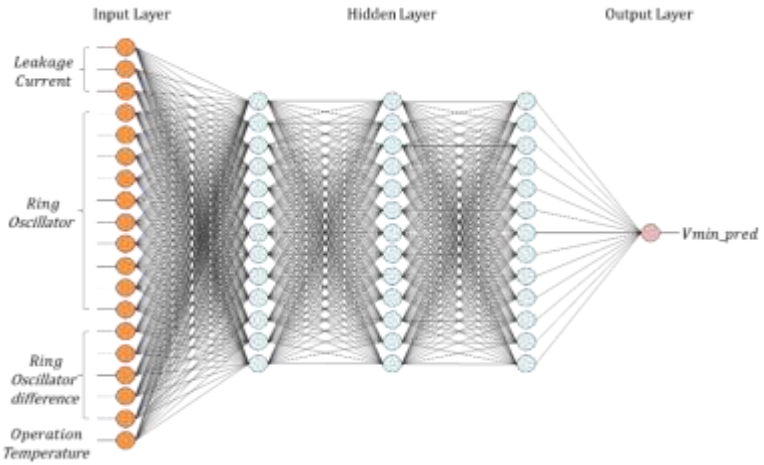


Fig. 2: Fully connected DNN with 3-hidden layers and 1-output layer.
19-characteristic elements are used as input. Each hidden layer has 13 neurons (2/3 of input layer nodes [7]).

3. Experiment

3.1. Setup of Experiments

We verified how accurately the V_{min} prediction was performed. We also experimented to verify how much the over-supplied voltage was improved. We selected 8nm SoC products in mass production and experimented with about 1500 APs that were initially produced. This may be a small number of datasets to build the model, but this experiment can be meaningful because it is a study of the possibilities and trends. Based on the results of this study, we will be conducting additional experiments to optimize the model with more data to be accumulated in the future.

$$normalized\ Voltage = \frac{V}{V_{max\_domin}} \tag{1}$$

$$RMSE_{calculated} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_{\min} - V_{min\_measured})^2} \tag{2}$$

$$normalized\ RMSE = \frac{RMSE_{calculated}}{Average\ of\ all\ RMSE} \tag{3}$$

In the company's security policy, the normalized value was used instead of the actual value. The normalized voltage is the ratio of the actual measured or predicted voltage divided by the maximum voltage of the power domain as shown in (1). The LC and RO values are divided by the mean of each normal distribution. The Root Mean Square Error (RMSE) is calculated by using the difference between V_{min} of VB or prediction model and $V_{min\_measured}$ as shown in equation (2). Equation (3) shows the normalized value of $RMSE_{calculated}$ divided by the average value of all RMSEs.

Experiments were conducted on three representative frequencies of high, mid, and low. We did not use the actual operating frequency value, but the highest frequency was denoted by *L0* and the lowest frequency was denoted by *L2*. We experimented on three of various power domains and named domains *A*, *B* and *C* without using real names.

The prediction accuracy of the model was evaluated as the RMSE and Pearson correlation coefficient. Experiments were performed by comparing $V_{min\_measured}$ of the sampling AP with the predicted output voltage ($V_{min\_predicted}$) of the proposed model. RMSE is a function mainly used to check the error between the actual value and the predicted value. The Pearson correlation coefficient is used mainly to analyze the correlation with focus on the linear relationship of two variable groups. It has a value between -1 and 1, and a positive coefficient of 1 means a perfect positive linear correlation, and a negative coefficient of -1 means a negative linear correlation. 0 means that there is no linear correlation between the two variables. In this experiment, the closer the correlation coefficient is to 1, the smaller the difference between $V_{min\_measured}$ and $V_{min\_predicted}$.

Fig. 3 shows the process of loss change of training dataset and validation dataset when training *L0* of power domain *C* using the proposed model. Other power domains configured the experimental environment in the same way and the loss graph was slightly different from each power domain frequency level.

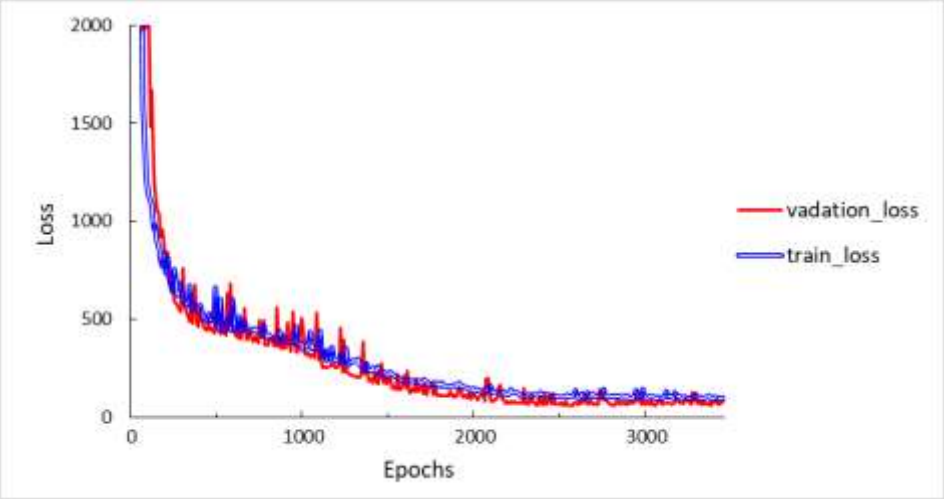


Fig. 3: Change of the MSE loss value during the model training.

To evaluate the accuracy and excellence of our model, we compared the results of ElasticNet, one of the most used linear regression models [11], and our model. For further analysis of the experiment, we also analyzed the results of 2Input_ElasticNet, which inputs only main RO (RO_{domain}) and LC (LC_{domain}) data of each power domain, which are same inputs used in the traditional method of AP VB.

3.2. Experiments Result

Table 1: Normalized RMSE by Equation (3).

Power Domain	Level	**OURS**	ElasticNet	2Input_ ElasticNet	VB
A	L0	**0.547**	0.694	0.982	3.349
	L1	**0.356**	0.356	0.557	1.270
	L2	**0.388**	0.461	0.822	1.230
B	L0	**0.692**	0.853	0.986	1.912
	L1	**0.334**	0.368	0.528	1.491
	L2	**0.357**	0.345	0.687	1.569
C	L0	**0.530**	0.776	1.020	4.900
	L1	**0.380**	0.392	1.122	2.291
	L2	**0.486**	0.499	1.382	1.086

Table 1 shows the comparison of normalized RMSE values for each model. The RMSE value of OURS was much lower than the traditional method, VB. The results of VB mean that the difference between $V_{min\_measured}$ and $V_{min\_predicted}$ is larger than the other models because the traditional method applies the same voltage to each group. Compared with OURS and ElasticNet, we found that there was a greater gain at the high-frequency level (*L0*) of all power domains than the other levels. At most levels, the results of OURS were better.

Table 2: Accuracy of models.

		±1 STEP Accuracy (%)			±2 STEP Accuracy (%)		
Power Domain	Level	**OURS**	ElasticNet	2Input_ ElasticNet	**OURS**	ElasticNet	2Input_ ElasticNet
A	L0	**82.77**	70.51	50.00	**97.44**	89.74	73.08
	L1	**97.59**	95.18	77.11	**100.00**	100.00	97.59
	L2	**93.18**	87.50	59.09	**98.86**	96.59	82.95
B	L0	**68.32**	59.41	55.45	**89.11**	84.16	83.17
	L1	**95.16**	94.12	82.35	**100.00**	100.00	92.65
	L2	**96.88**	98.44	70.31	**100.00**	100.00	87.50
C	L0	**80.77**	52.88	50.48	**97.12**	87.50	67.62
	L1	**95.89**	89.04	55.41	**100.00**	100.00	78.38
	L2	**95.00**	90.00	27.05	**97.50**	97.50	57.50
Average		**89.43**	81.90	58.63	**97.78**	95.05	80.05

Table 2 shows the accuracy of the prediction model by *STEP*. 1 *STEP* is the smallest unit voltage of Power Management Integrated Circuits (PMIC) that apply the actual voltage to the AP. ±1 *STEP* Accuracy table shows the accuracy of whether $V_{min\_predicted}$ is included in the $V_{min\_mearsured} \pm 1$ *STEP* range. Analyzing the ±1 *STEP* accuracy to evaluate how close $V_{min\_predicted}$ is to $V_{min\_measured}$ shows that the accuracy of OURS at most levels was higher than other models. Furthermore, the above described RMSE reduction effect of *L0* was confirmed again in accuracy results. The results of comparing our prediction

model with ElasticNet, linear regression model, are as follows. While accuracy was improved average 3.24% in the *L1* and *L2*, in case of *L0* where the AP operates at the fastest speed, it was improved average 16.11%, up to 27.88%. From a power domain perspective, average accuracy of power domain *A* and *B* increased by 6.54% and 2.80%, respectively, and power domain *C* increased by average 13.24%. In summary, it can be said that our DNN model describes more complicated formulas well than other models.

Table 3: Pearson correlation coefficient.
The closer to 1 the coefficient is the positive linear correlation.

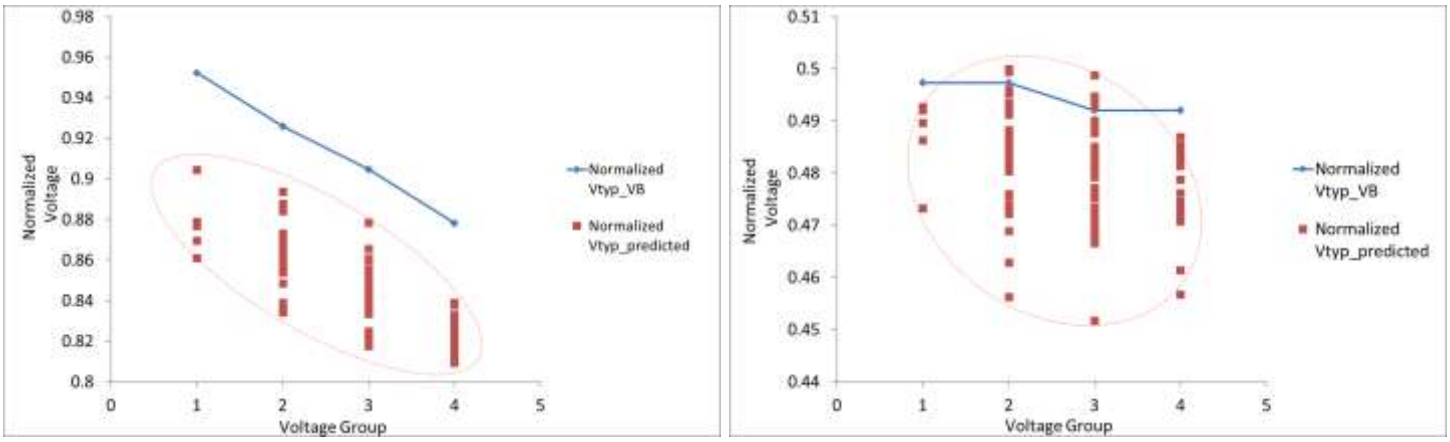
Power Domain	Level	**OURS**	ElasticNet	2Input_ ElasticNet	VB
A	L0	**0.948**	0.924	0.846	0.739
	L1	**0.913**	0.916	0.760	0.743
	L2	**0.900**	0.858	0.389	0.341
B	L0	**0.951**	0.924	0.895	0.892
	L1	**0.925**	0.925	0.821	0.795
	L2	**0.890**	0.895	0.500	0.543
C	L0	**0.947**	0.891	0.827	0.773
	L1	**0.933**	0.932	0.583	0.673
	L2	**0.773**	0.728	-0.167	0.296

Table 3 shows the Pearson correlation coefficient which model has the highest positive correlation with $V_{min\_measured}$. We confirmed that the Pearson value of other models was higher than VB. In particular, our model had the closest value to 1 at most levels, which is the best representation of the relationship between $V_{min\_measured}$ and $V_{min\_predicted}$.

Table 4: Normalized typical voltage (V_{typ}) of VB and our model by Equation (1).

Power Domain	Level	V_{typ_VB}	**$V_{typ\_predicted}$**	**Enhancement**
A	L0	0.909	**0.848**	**6.76%**
	L1	0.663	**0.646**	**2.58%**
	L2	0.494	**0.481**	**2.74%**
B	L0	0.899	**0.863**	**4.00%**
	L1	0.658	**0.642**	**2.44%**
	L2	0.507	**0.486**	**4.03%**
C	L0	0.897	**0.838**	**6.68%**
	L1	0.634	**0.607**	**4.32%**
	L2	0.476	**0.466**	**2.10%**

Considering the difference between the actual environment in which the mobile device is used and the idle manufacturing environment in which V_{min} is measured, a slightly higher voltage than V_{min} is applied in actual usage environment. This is called the typical voltage (V_{typ}). Table 4 shows how much the voltage converted from $V_{min\_predicted}$ to the predicted typical voltage ($V_{typ\_predicted}$) is lower than the typical voltage using VB (V_{typ_VB}). We calculated the voltage averages of all APs corresponding to each power domain frequency level. There was an average 3.96% voltage reduction compared to the traditional method for all power domains. Especially, at the high-frequency level (*L0)*, the voltage reduction was up to 6.76%, which is larger than other levels.



(a) Power Domain *A, L0* (b) Power Domain *A, L2*

Fig. 4: Normalized voltage distribution.

Fig. 4 shows the voltage distribution of $V_{typ\_predicted}$ after optimizing the AP voltage which used to apply the V_{typ_VB}. In the case of the power domain *A L0* in Fig. 4 (a), voltage group 1 APs with 0.952 V_{typ_VB} determined by VB were reduced to range of 0.904 to 0.861 $V_{typ\_predicted}$ by using our model. There was a clear voltage difference between V_{typ_VB} and $V_{typ\_predicted}$ where the voltage reduction was 4.80%. In addition, as shown in the domain *A L2* graph in Fig. 4 (b), most of the APs had voltage reduction compared to V_{typ_VB}.

Based on the above experimental results, we have confirmed that the proposed V_{min} prediction model can be an effective way to reduce the over-supplied voltage.

3.3. Considerations in Manufacturing Process

There are additional considerations. When $V_{min\_measured}$ and $V_{min\_predicted}$ are compared, if $V_{min\_predicted}$ is significantly lower than $V_{min\_measured}$, there may be a case that the voltage required for the AP operation is insufficient and malfunction may occur. Therefore, post-treatment may be required in manufacturing process.

To avoid above problem, the process test should proceed with a slight overvoltage condition so that the yield and evaluation time will not be affected. The condition can be determined by referring to the accuracy table. For example, as shown in Table 2, prediction accuracy of power domain *A L0* is 97.44% in ± 2 *STEP* range, and if APs are tested in +2 *STEP* voltage higher than $V_{min\_predicted}$, 98.72% will not have problem in the manufacturing process. For the remaining 1.28% below -2 *STEP*, it requires retest after additional voltage *STEP* rise. This retest portion is less than what we can get with traditional VB method.

In this way, *STEP* level should be determined in consideration of the power domain and the frequency level, and we should make the best choice to get the maximum voltage gain with a minimum voltage rise.

4. Conclusion

In this paper, we proposed a model to predict V_{min} in the AP mass production process. We analyzed the relationship with many elements related to V_{min} and used it as an input of the model. By using the DNN to analyze the complex equations well and produce the best results, the accuracy of the model was improved and the voltage gain was up to 6.76%. In addition, when comparing the results of ElasticNet, a linear regression model, and ours, it was confirmed that the accuracy of the our proposed model at a high-frequency level (*L0)* was much higher. Therefore, the proposed model can be an effective solution to accurately predict V_{min} and reduce over-supplied voltage.

References

[1] G. Singla, G. Kaur, A. K. Unver, and U. Y. Ogras, "Predictive dynamic thermal and power management for heterogeneous mobile platforms," in *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, 2015, pp. 960–965.

[2] S. Lichtensteiger and J. P. Bickford, "Using selective voltage binning to maximize yield," *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 4, pp. 436–441, 2013.

[3] V. Zolotov, C. Visweswariah, and J. Xiong, "Voltage binning under process variation," in *Proceedings of the 2009 International Conference on Computer-Aided Design*, 2009, pp. 425–432.

[4] R. Shen, S. X.-D. Tan, and X.-X. Liu, "A new voltage binning technique for yield improvement based on graph theory," in *Thirteenth International Symposium on Quality Electronic Design (ISQED)*, 2012, pp. 243–248.

[5] L. Croce and S. T. I. Universita, "Measuring the effects of process variations on circuit performance by means of digitally-controllable ring oscillators," *Int. Conf. Microelectron. Test Struct.*, pp. 214–217, 2003.

[6] H. Sultan, S. Varshney, and S. R. Sarangi, "Is Leakage Power a Linear Function of Temperature?" *arXiv Prepr. arXiv1809.03147*, 2018.

[7] S. Karsoliya, "Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture," *Int. J. Eng. Trends Technol.*, vol. 3, no. 6, pp. 714–717, 2012.

[8] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in neural information processing systems*, 2017, pp. 971–980.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[10] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the trade*, Springer, 1998, pp. 55–69.

[11] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B (statistical Methodology).*, vol. 67, no. 2, pp. 301–320, 2005.

[12] M. K. Mandai and B. C. Sarkar, "Ring oscillators: Characteristics and applications," *Indian J. Pure Appl. Phys.*, 2010.

[13] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017.

[14] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, 2016.