Neural Approach for Modeling and Optimizing Si-MOSFET Manufacturing

HYUN-CHUL CHOI\(^1\), HYEOK YUN\(^2\), (Graduate Student Member, IEEE), JUN-SIK YOON\(^2\), (Member, IEEE), AND ROCK-HYUN BAEK\(^2\), (Member, IEEE)

\(^1\)Department of Electronic Engineering, Yeungnam University, Gyeongsan 38541, South Korea
\(^2\)Department of Electrical Engineering, Pohang University of Science and Technology, Pohang 37673, South Korea

Corresponding author: Rock-Hyun Baek (rh.baek@postech.ac.kr)

This work was supported in part by the Ministry of Trade, Industry and Energy under Grant 10080617, in part by the Korea Semiconductor Research Consortium Support Program for the Development of the Future Semiconductor Device, in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning under Grant NRF-2018R1C1B6004056 and Grant NRF-2020R1A4A4079777, in part by the Semiconductor Industry Collaborative Project between Pohang University of Science and Technology (POSTECH) and SK Hynix Inc., and in part by IC Design.

ABSTRACT An optimal design of semiconductor device and its process uniformity are critical factors affecting desired figure-of-merits as well as reducing fabrication cost of fixing possible malfunctioning in semiconductor manufacturing. Two main tasks in optimal device design for semiconductor manufacturing, i.e., parameter optimization and modeling, have been typically used either to characterize the devices by understanding how each parameter affects the device performances or to calibrate the parameters for SPICE circuit simulation. However, there still remains limitations in describing the relationship between all manufacturing parameters and figure-of-merits using several simple equations human experts can utilize. Even with the best model currently available, the optimal design of semiconductor device heavily relies on experiences of human experts and deals with time-consuming ad-hoc trials and non-holistic approaches. In this paper, we propose a new approach for data-based accurate electrical modeling of transistor, which is the most fundamental unit device of semiconductor, and fast optimization of its manufacturing parameters. Instead of the previous analytic approaches, finding finite equations derived from semiconductor physics, we utilize machine learning technique and neural networks to find appropriate modeling functions from data pairs of parameters and figure-of-merits. And for given desired figure-of-merits, we find optimal manufacturing parameters in holistic manner by using the learned functions of neural networks and fast gradient-based optimization method. Experimental results show that our neural-network-based-model directly estimate figure-of-merits with competitive accuracy and that our holistic optimization technique accurately and rapidly adapts the manufacturing parameters to meet desired figure-of-merits.

INDEX TERMS Transistor, semiconductor manufacturing, holistic parameter optimization, machine learning, neural networks, gradient descent method.

I. INTRODUCTION Transistor process is the most basic and core technology in semiconductor industry. Thus, developing a transistor that is high-performing and low-power-consuming is of a great importance. Typically finding optimal manufacturing parameters for high performing and low power consuming transistor requires a repeated wafers fabrication with hundreds of unit process steps to measure its electrical performances and a feedback loop to unit process of manufacturing. This procedure dramatically increases research and development cost because it can take up to several weeks. The duration of the procedure has been extended with the shrinkage of technology node.

To reduce the time consuming trials, multiple equation-based models have been suggested in analytic or numerical forms either to physically understand device characteristics [1] or to model I-V and C-V through tens of fitting parameters for SPICE circuit simulation [2], [3]. However, these modeling equations do not explain the relationship...
between manufacturing parameters and figure-of-merits at a time. There is no simple equation to delineate the relationship and ultimately to meet given desired figure-of-merits by understanding their correlations. In addition, even with those equations, manufacturing parameter optimization heavily relies on experiences of human experts and has been dealt with ad-hoc trials and non-holistic approaches. For instance, human experts design a transistor to optimize DC performance firstly by maximizing on/off current ratio from I-V characteristics and AC performance secondly by adjusting capacitance from C-V characteristics because of the antagonistic effects between I-V and C-V characteristics on DC/AC performances of transistor when its structure changes.

In this paper, we aim to overcome time-consuming and non-holistic optimization of transistor manufacturing parameters. For this end, we propose a new approach for both data-based precise transistor modeling and fast optimization of transistor manufacturing parameters. Instead of the previous analytic approaches in finding finite equations derived from semiconductor physics, we utilize artificial intelligence (AI), namely neural networks [4], [5], and machine learning technique, to automatically find appropriate modeling functions from a number of data pairs of input manufacturing parameters and output figure-of-merits. Subsequently, we find optimal manufacturing parameters for given desired figure-of-merits in automatic and holistic manner via gradient-based optimization technique and the learned modeling functions.

AI has been previously applied for monitoring the excursion, i.e., out of boundary figure-of-merits in unit process of transistor manufacturing [6], but there is no AI-based study on analyzing the relationship between overall manufacturing procedure and figure-of-merits of transistor. Our holistic approach simultaneously optimizes the entire manufacturing procedure in the whole parameter space while conventional sequential optimization for each unit process results in a sub-optimal transistor due to its limited parameter-searching ability. The contributions of this work are as follows:

1) Physics-free modeling and high compatibility: In semiconductor field, different models with regards to material, scale, structure, and usage of transistor are conventionally required. However, our AI model replaces multiple models with one artificial neural network model, which is independent of those physical characteristics and able to be applied to various transistor only with their input/output data pairs. In the aspect of variability, our method demonstrates the correlation between manufacturing parameters and device figure-of-merits from a number of data even for nanoscale devices suffering from physical ambiguity.

2) Fast modeling and optimization: It takes several hours for excellent human expert to compactly model a given transistor. In contrast, our AI method takes only several tens of seconds to complete the same task. Similarly, our AI method takes only a couple of minutes to find optimal manufacturing parameters for given device figure-of-merits.

3) Fast prediction without accuracy penalty: To speed up TCAD simulation in transistor design, less physical parameters and more empirical parameters are necessary. This results in low accuracy of simulation result. However, our AI method uses a compact neural network for fast prediction and guarantees high accuracy as the number of data increases.

4) Holistic and flexible optimization: Our method finds manufacturing parameters to optimize AC performance of transistor by simultaneously considering I-V and C-V parameters. In addition, our method can also selectively optimize the performance suitable for application of transistor, which is useful for flexible design of transistor for System-On-Chip application.

In the remained of this paper, we will describe the details of our AI-based method in sec.II, experimentally prove the superiority of our method in sec.III, and conclude this work in sec.IV.

II. METHOD

A. TRANSISTOR STRUCTURE AND PARAMETERS

Since the purpose of this work is not to find a new low-scaled devices but to optimize transistor and to analyze its variation, we designed 32-nm node high-K metal gate transistor, a 2-D planar MOSFET which is easily dealt with in data acquisition compared to 3-D FinFET in the current 7 nm technology node, and simulated electrical characteristics using Sentaurus TCAD simulator [7]. The transistor structure was designed using structure editor with parameters based on Intel 32-nm node [8]. Values and ranges of structure parameters are shown in fig.1 and summarized in table 1. All doping profiles were assumed as the Gaussian distribution. The junction gradient lengths for source/drain ($L_{adj}$) and halo implants ($L_{halo}$) indicate the lengths where the source/drain ($N_{sd}$) and halo ($N_{halo}$) implants are reduced by 10 times, respectively.

B. NEURAL-NETWORK-BASED MODELING FOR POWER AND DELAY PREDICTION

In order to avoid very complex mathematical equations of semiconductor physics and to make a general model for
multiple devices, we utilize one of the recently popularized hyperparametric models, namely artificial neural networks (ANN) [4], [5], and machine learning techniques to model functions that output figure-of-merits of transistor with the given input vector. The unit structure of ANN is perceptron [9], which models after the biological function of neuronal synapse as a linear combination of inputs followed by a non-linear activation function. ANN with several layers of multiple perceptrons is called fully connected multi-layer perceptron (MLP) [10], [11]. The advantage of MLP is in making a flexible model for new device by simply changing the number of input and output perceptrons, regarding to the dimension of manufacturing parameters and output figure-of-merits. In this work, we use two of 3-layered MLP (input-hidden-output layers) as shown in fig.2 because the structure is simple enough to control non-linearity function by changing the number of perceptrons in the hidden layer.

By using these two MLPs, we model two functions of transistor, \( F_{\text{power}} \) (1) and \( F_{\text{delay}} \) (2), which respectively output log-scaled power \( \log(O_{\text{power}}) \) and delay \( \log(O_{\text{delay}}) \) of a device from given input log-scaled vector \( \log(I_p) \) of manufacturing parameters \( p \), where power and delay, or power delay products (PDP) are figure-of-merits for digital logic applications [12].

\[
\begin{align*}
\log(O_{\text{power}}) &= F_{\text{power}}(\log(I_p)), \\
\log(O_{\text{delay}}) &= F_{\text{delay}}(\log(I_p)).
\end{align*}
\]

Here, we used \( \log \) values in input and output of \( F_y \) for \( y = \text{power or delay} \) to avoid network training failure caused by large scale differences between input and output values.

We follow the data flow and operations presented in fig.3 for MLP training. We use mean squared error (MSE), \( L_y \) (3), between model output \( \log(O_y^i) \) and desired output \( \log(D_y^i) \) of \( i \)-th training sample \( \langle I_p^i, D_y^i \rangle \) for \( i = 1..N \) as the only loss function. Each MLP for \( y = \text{power or delay} \), is trained to minimize this loss function \( L_y \) (3) by backpropagation procedure [13], which updates the hyperparameters of MLP from initial random values. Minimizing \( L_y \) in \( \log \) scale is identical to minimizing log-scaled ratio, \( |\log(O_y^i) - \log(D_y^i)| \), which considers both scale-sensitive ratio and log-scaled error symmetric regarding to predicted and target values.

\[
L_y = \frac{1}{N} \sum_{i=1}^{N} |\log(O_y^i) - \log(D_y^i)|^2, \quad y = \text{power or delay}.
\]

One could suggest using a single MLP, which outputs both power and delay from given input manufacturing parameters. Such an approach is also possible and, in actuality, more efficient in the size of neural networks because two different outputs share the same hidden layer. However, in training procedure, a multi-objective loss function \( L_m \) of two objectives (3) corresponding to power and delay should be defined.
and usually has a form of weighted summation (4) of $L_{power}$ and $L_{delay}$ with ambiguous weights $w_{power}$ and $w_{delay}$.

$$L_m = w_{power} \cdot L_{power} + w_{delay} \cdot L_{delay}. \quad (4)$$

Since training a single MLP by minimizing $L_m$ may result in overly fitted $F_y$ with a larger weight $w_y$ or less optimized $F_y$ with a small weight $w_y$, we need to repeat MLP training process with varying weights to get satisfactory values of $L_{power}$ and $L_{delay}$ to determine appropriate values for the weights. In contrast, when using two MLPs for two outputs as in our approach (fig.2), it is unnecessary to find appropriate weights for multiple loss functions and each MLP is independently trained with its own loss function $L_y$ (3).

C. AUTOMATIC OPTIMIZATION OF MANUFACTURING PARAMETERS WITHIN CONTROLLABLE RANGE

Once two models of power and delay are ready to use, we can not only forwardly predict power or delay from given manufacturing parameters but also backwardly optimize manufacturing parameters for a better performance of transistor. PDP is a commonly used measurement of transistor performance and should be as small as possible. Since our neural networks are designed to output log-scaled power and delay (fig.2), log-scaled PDP can be represented as the summation of two outputs of our networks (5).

$$log(PDP(x)) = log(O_{power}(x) \cdot O_{delay}(x))$$

$$= log(O_{power}(x)) + log(O_{delay}(x))$$

$$= F_{power}(x) + F_{delay}(x),$$

$$x = log(I_p). \quad (5)$$

As shown in fig.4, we find the optimal set of manufacturing parameters, which minimizes this log-scaled PDP value by utilizing our trained neural networks and the gradient descent method [14], [15]. Gradient descent method operates as described here: initialize randomized manufacturing parameters, $I_p^0$, and learning rate $\gamma$ are set as a preliminary step. Then, as the first step, the gradient of log-scaled PDP with respect to manufacturing parameter $G(x)$ (6) is calculated with the initial manufacturing parameters ($x = log(I_p^0)$). As the second step, the initial manufacturing parameters are updated to move to the negative direction of the gradient with a certain learning rate $\gamma$ as (7) at time stamp $t = 0$. Those two steps are repeated with the updated manufacturing parameters as the time stamp $t$ increases until PDP value meets a termination condition (8), i.e., PDP less than a certain threshold value $T_{PDP}$ or vanishing gradient or no variation in manufacturing parameters.

$$G(x) = \frac{d(log(PDP(x)))}{dx}$$

$$= \frac{dF_{power}(x)}{dx} + \frac{dF_{delay}(x)}{dx},$$

$$x^{t+1} = x^t - \gamma \cdot G(x^t), \quad x^t = log(I_p^t), \quad (7)$$

$$PDP(x^t) \leq T_{PDP} \text{ or } G(x) \leq T_G \text{ or } dx \leq T_{dx}. \quad (8)$$

Here, the manufacturing parameters obtained from this gradient descent method are not guaranteed to be in the controllable range of real semiconductor manufacturing because the method may find an optimal solution along the entire hypersurface of our trained neural networks beyond the training data space, without any constraint of the acceptable range of solution in reality. Thus setting a controllable range of manufacturing parameters in semiconductor industry is critical because of both the limitation of manufacturing hardware and the preferred parameter setting. To provide a controllable optimal solution, as shown in fig.4, we use a limiter function [16] $g(z_p)$ (9) and optimize latent variables $z_p$ instead of manufacturing parameters $I_p$ themselves.

$$g(z_p) = \frac{1}{1 + e^{-z_p}}. \quad (10)$$

Because the output range is from 0 to 1 ($\{g(z_p) \in [0, 1]\}$), $g(z_p)$ becomes $log(I_p^{max})$ when $z_p$ goes to $+\infty$ and $log(I_p^{min})$ when $z_p$ goes to $-\infty$. Therefore, to find optimal manufacturing parameters in the controllable range $[I_p^{min}, I_p^{max}]$, we optimize $z_p$ without range constraint and calculate $I_p = g(z_p)$.

For numerical efficiency, instead of calculating complex analytic gradient (6) of neural network function, we use numerical gradient [18] $\tilde{G}$ (11) in our process optimization, where two feed-forward operations of (5) with the manufacturing parameters $x + \frac{\tau}{2}$ and $x - \frac{\tau}{2}$ and one division by $\tau$ $\approx 0$ are needed. Thus, updating equation of latent variables (12)
is also used instead of the equation for manufacturing parameters (7).

\[ \hat{G}(x) = \frac{\log(PDP \circ g(x + \frac{\tau}{2})) - \log(PDP \circ g(x - \frac{\tau}{2}))}{\tau}, \quad (11) \]

\[ x^{t+1} = x^t - \gamma \cdot \hat{G}(x^t), \quad x^0 = x^i, \quad (12) \]

### III. EXPERIMENTS

#### A. EXPERIMENTAL SETUP

For experimental data set, we gathered 1,000 samples of random manufacturing parameters in the reference range between min and max values specified in table 1 and simulated 1,000 samples of \( I_d - V_g \) and \( C_{gg} - V_g \) characteristics of the devices as shown in fig.5 by using Sentaurus TCAD simulator. Before the electrical simulation, we obtained the geometrical parameters such as gate length and equivalent oxide thickness announced in [8] and [19]. Then, we varied the doping profiles from source/drain to channel to calibrate subthreshold characteristics between simulated and Intel measured data [8]. Simulated I-V characteristics correspond to 32-nm node HP (High Performance) and LOP (Low Operating Power) targets [19] about the range of the threshold voltage and on/off current ratio which are related to electrostatic characteristics like subthreshold swing and drain-induced barrier lowering. To obtain power (14) and delay (15), effective current \( I_{eff} \) is extracted from \( I_H \) and \( I_L \) (13) using [21]. We extracted \( I_H \) at gate voltage, \( V_g = V_{dd} \) and drain voltage, \( V_d = 0.5V_{dd} \) and \( I_L \) at \( V_g = 0.5V_{dd} \) and \( V_d = V_{dd} \).

\[ I_{eff} = (I_H - I_L)/\ln \left( \frac{I_H}{I_L} \right), \quad (13) \]

\[ \text{power} = I_{eff} \cdot V_{dd} + I_{eff} \cdot V_{dd}, \quad (14) \]

\[ \text{delay} = \frac{C_{gg} \cdot V_{dd}}{2 \cdot I_{eff}}. \quad (15) \]

Among these 1,000 samples of \{manufacturing parameters, delay and power\} pairs, 850 samples were used to train the models of delay and power, the other 150 samples were used to test the delay and power prediction performance of the trained models.

For TCAD simulation, drift-diffusion transport model was calculated self-consistently with Poisson and the carrier continuity equations. Density-gradient model was applied to consider the quantum confinement of carriers. Multivalley modified local-density approximation was used with six-band and two-band k.p band structures for holes and electrons, respectively. Canali model [20] for carrier velocity saturation at high electric field, Masetti model [22] to consider doping-dependent mobility, and Lombardi model [23] to consider mobility degradation at Si-SiO\(_2\) interface were used. Doping-dependent Shockley-Read-Hall [24] and Auger [25] generation-recombination models were used along with Hurkx band-to-band tunneling model [26] to consider the gate-induced drain leakage. Operation voltage is fixed to 1.0 V, and off-state current is fixed at 1 nA for the averaged device.

As a pre-processing of data set, we took logarithm of each element of the 850 train samples and then normalized to have (mean, standard deviation) = (0, 1) across samples. This pre-processing of eliminating bias and deviation differences between data helps multi-layered perceptron (MLP) to learn a function in the restricted range of domain and is known to enhance training performance [27]. We performed the same pre-processing for the 150 test samples by using the mean and standard deviation of the train samples.

All experiments were operated in MATLAB console on a laptop computer with Intel i7-8550U (Quadcore, 1.8GHz) CPU and 16 GByte RAM.

#### B. PERFORMANCE EVALUATION OF POWER AND DELAY MODELING

We built two neural network models, one for predicting delay and the other for power of transistor, (fig.2) as 3-layered MLPs of 19 nodes for input layer, 7 nodes for hidden layer, and 1 node for output layer, where the number of nodes in the hidden layer were experimentally determined as the smallest number to learn regression between manufacturing
parameters and delay or power without overfitting. In details, the two models were trained, one to minimize power loss of (3) when \( y = \text{power} \), the other to minimize delay loss of (3) when \( y = \text{delay} \). During the training phase, randomly selected 700 samples out of 850 samples (termed training set) were used to update neural network parameters using Levenberg-Marquardt (LM) optimizer. The other 150 samples (termed validation set) were used find the stopping point for training iteration based on the generalization performance of the trained neural networks. In our experiments, we stopped the iteration of LM optimizer when the loss \( L_y \) for validation set continuously remained or increased during the previous 6 epochs. We repeated this training process for varying number of nodes in the hidden node and used the smallest number of nodes that produce small training error for training set and near-zero difference between errors of training and validation sets.

Since the performance of a trained network differs based on its initial random parameter values, we repeated this training procedure with the same training and validation sets but with different initial randomized values of network parameters. And then, we took the best network which resulted the least error for training and validation sets. To find an appropriate number of repetition with a stable learning performance, we measured average MSEs and their variations of delay (\( F_{\text{delay}} \)) and power (\( F_{\text{power}} \)) models from 10 trials of training with each repeat number. As shown in fig.6, average loss and its variation decrease as the repeat number increases but no critical improvement was resulted after 40 repeats. Therefore, considering moderate training time, we concluded to use the best models from 40 trained networks.

For the trained models, we firstly analyzed the learning ability and generalization performance of our neural models by monitoring the Loss (3) reduction during training process. Figure 7 shows loss (MSE error) reductions of the normalized data sets (sec.III-A) during training process. For both training (blue lines) and validation (green lines) sets, the loss gradually decreased as the training epoch continued until optimizer stopped (fig.7a and fig.7b). Even for test data set (red lines), which was not shown in training procedure, the same loss to the validation set was achieved. Figure 8 shows output regression lines of our two trained neural models. The regression plots of delay model (fig.8a) and power model (fig.8b) show almost ideal linear lines with R values very close to 1.0 for all kinds of data sets. Based on these results, we concluded that our neural models learned the deterministic equations of delay and power for given manufacturing parameters and that our neural models have a generalization performance to predict delay or power for unseen manufacturing parameters in the same perturbation range of training data.

Secondly, to evaluate the modeling accuracy of our neural approach, we compared the predicted delays and powers of our neural models to the simulated delays and powers (ground truth) for the 150 test samples in sec.III-A. As shown in fig.9, the differences between predicted and simulated values are...
very small (left panel of fig.9), less than 5% of the simulated values (right panel of fig.9), and this prediction accuracy is good enough to apply our neural models to predict delay and power in a real manufacturing process.

Our neural modeling took 80 seconds to train two neural networks 40 times. Additionally, compared to TCAD simulation that takes 3 min, our method takes less than 1 ms to perform a forward prediction of delay and power for a certain manufacturing parameter.

C. PERFORMANCE EVALUATION OF AUTOMATIC MANUFACTURING PARAMETER OPTIMIZATION

In the optimization process represented in fig.4, we fixed our trained neural models of delay and power in sec.III-B and used them as $F_{\text{delay}}$ and $F_{\text{power}}$ respectively. And then, we generated random initial values of manufacturing parameters and updated them by gradient descent method as described in sec.II-C with numerical gradient (11), update equation (12), and objective function (5). Random initial values of manufacturing parameters were generated under the uniform distribution in a reasonable range, i.e., from $I_{\text{min}}^p$ to $I_{\text{max}}^p$ of (9) where $I_{\text{min}}^p = 0.95 \times \text{reference value in table 1}$ (-5%) and $I_{\text{max}}^p = 1.05 \times \text{reference value in table 1}$ (+5%). The small deviation $\tau = 2 \times 10^{-8}$ was used for calculating accurate numerical gradient (11) and the learning rate $\gamma = 10^{-3}$ was used for update equation (12) for fast convergence. $T_{\text{PDP}} = 0$, $T_G = 10^{-5}$, and $T_{dx} = 10^{-5}$ were used for termination condition (8).

Since gradient descent method does not always guarantee global optimum from any initial value, we repeated this optimization process with different initial values and took the best result as the optimal manufacturing parameters. To find the necessary number of repetition of optimization process to obtain stable optimal result, we firstly measured the best PDP value for each repeat number and secondly calculated mean and standard deviation from 10 trials. Figure 10a and the second row of table 2 show the measured values where predicted optimum and simulated optimum represent PDP values corresponding to the optimized manufacturing parameters from our neural model and from Sentaurus TCAD simulator, respectively. We find that the predicted PDP is saturated at around $0.3066 \times 10^{-15}$ when repeat number is $\geq 10$ and its deviation diminished to $0.0002 \times 10^{-18}$. Even without any repeat (when repeat number is 1), the resulted PDP value ($0.3071 \times 10^{-15}$ with standard deviation of $0.9429 \times 10^{-19}$) is in close proximity to the saturated optimal value. Compared to the ground truth (simulated optimum), the predicted optimum has slightly lower value. This difference is expected and caused by the model prediction error shown in fig.9 but the
When repeat number is 20, delay (fig.10b) and power (fig.10c) converged to the best predicted optimum, 0.1685 × 10⁻¹¹ (third row of table 2) and 0.1819 × 10⁻³ (fourth row of table 2) respectively. Compared to the simulated optima of delay and power, the predicted optima have very small difference within ±5% bounds (fig.10b and fig.10c) of the simulated optima. Based on these results, we decided to repeat optimization process 40 times (20 + marginal 20) to achieve the lowest PDP performance with stable delay and power.

Our optimized PDP and its corresponding delay and power are compared to the values of 150 test samples (III-A) in fig.11. As shown in fig.11a, our optimized PDP (green line) is lower than the minimum PDP of the test samples. However, the minimum PDP was not resulted with the minima of both delay and power. Instead, our optimized manufacturing parameters have relatively high delay (fig.11b) and relatively low power (fig.11c). This result is partially consistent with the previous optimizing technique of human expert where human expert firstly minimizes power (DC characteristics) by maintaining good short channel characteristics and then tries to minimize delay (AC characteristics) under the low power condition. However, our neural optimization result differs in that it does not insist on the lowest power and delay; rather our approach resulted in a slightly higher power, yet achieved the lowest PDP with a certain delay.

Our optimized PDP and its corresponding delay and power are compared to the values of 150 test samples (III-A) in fig.11. As shown in fig.11a, our optimized PDP (green line) is lower than the minimum PDP of the test samples. However, the minimum PDP was not resulted with the minima of both delay and power. Instead, our optimized manufacturing parameters have relatively high delay (fig.11b) and relatively low power (fig.11c). This result is partially consistent with the previous optimizing technique of human expert where human expert firstly minimizes power (DC characteristics) by maintaining good short channel characteristics and then tries to minimize delay (AC characteristics) under the low power condition. However, our neural optimization result differs in that it does not insist on the lowest power and delay; rather our approach resulted in a slightly higher power, yet achieved the lowest PDP with a certain delay.

Our optimized PDP and its corresponding delay and power are compared to the values of 150 test samples (III-A) in fig.11. As shown in fig.11a, our optimized PDP (green line) is lower than the minimum PDP of the test samples. However, the minimum PDP was not resulted with the minima of both delay and power. Instead, our optimized manufacturing parameters have relatively high delay (fig.11b) and relatively low power (fig.11c). This result is partially consistent with the previous optimizing technique of human expert where human expert firstly minimizes power (DC characteristics) by maintaining good short channel characteristics and then tries to minimize delay (AC characteristics) under the low power condition. However, our neural optimization result differs in that it does not insist on the lowest power and delay; rather our approach resulted in a slightly higher power, yet achieved the lowest PDP with a certain delay. Therefore, as an holistic optimization, our neural approach is promising in finding optimal manufacturing parameters for the lowest PDP while the commonly used sequential optimization in real manufacturing, which firstly sticks to the minimal power and secondly searches minimal delay, does not guarantee the lowest PDP. Benefits of our neural approach will be further proved by quantitative comparing with human expert and the previously used reference design in sec.III-E.

Table 3 presents the average and standard deviation values of each optimized manufacturing parameter when repeat number is 40. Each parameter converged to a certain optimal value (Avg) with a very small deviation (Std) of 10⁻⁴ or less scale of its corresponding Avg value (Std / Avg). Figure 12 shows the optimized transistor structure of the manufacturing parameters shown in table 3. Since our optimization method has no constraint of source/drain symmetry, the resulted optimal transistor has an asymmetric structure unlike commonly used symmetric structure [8]. Based on the table 2, this asymmetric structure achieved 17% lower PDP (0.3120 × 10⁻¹⁵, 12% higher delay (0.1699 × 10⁻¹¹), and 26% lower power (0.1836 × 10⁻³) than reference values, 0.3749 × 10⁻¹⁵, 0.1516 × 10⁻¹¹, and 0.2473 × 10⁻³ respectively, which were obtained using TCAD simulation with the reference structure [8] in table 1.

Although asymmetric structure is not common and beyond the present manufacturing ability, this optimization result shows that our neural approach can suggest theoretical direction for MOSFET structure as well as optimal parameter values for better performance beyond the conventional source/drain symmetric structure. Our optimization method

### TABLE 2. Optimized values vs. number of repeated optimization process with 5% limiter: values in parentheses are standard deviations corresponding to the above averaged values.

| Repeat number | PDP × 10⁻¹⁵ (× 10⁻¹⁸) | Delay × 10⁻¹¹ (× 10⁻¹³) | Power × 10⁻⁵ (× 10⁻⁷) |
|---------------|------------------------|--------------------------|------------------------|
| 10            | 0.3071 (0.9429)        | 0.1661 (0.2284)          | 0.1849 (0.2990)        |
| 20            | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1820 (0.00005)       |
| 30            | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1819 (0.00005)       |
| 40            | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1819 (0.00005)       |
| 50            | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1819 (0.00005)       |
| 60            | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1819 (0.00005)       |
| 70            | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1819 (0.00005)       |
| 80            | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1819 (0.00005)       |
| 90            | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1819 (0.00005)       |
| 100           | 0.3066 (0.00002)       | 0.1685 (0.00003)         | 0.1819 (0.00005)       |
FIGURE 10. Optimized values of PDP, delay, and power vs the number of repeated optimization process: predicted optimum represents the output value from our neural model corresponding to the optimized manufacturing parameters, simulated optimum represents the values from TCAD simulator, upper and lower bounds represents 5% higher and lower values respectively than the simulated optimum when repeat number is 20.

We can also find a set of optimal source/drain symmetric manufacturing parameters by applying our optimization technique of sec.II-C assuming the same parameter for source/drain. The symmetric structure optimization will be dealt with in sec.III-E.

D. CORRELATION AND SENSITIVITY ANALYSIS

We additionally analyzed the effects of each manufacturing parameter on PDP value in optimization process by using our trained neural models of sec.III-B and the optimized manufacturing parameters of sec.III-C. Each graph of fig.13, fig.14, and fig.15 shows transitions of figure-of-merits, i.e., PDP, delay, and power, respectively for each manufacturing parameter varying from Min to Max of table 1, while the other parameters are fixed at the optimal values of table 3. We calculated the differences between the minimum and maximum values of figure-of-merits as the sensitivity (dy) of parameter, i.e., how significantly a parameter affects the figure-of-merit. These graphs can be generated in a second from our neural models without using repeated TCAD simulations.

From these graphs, we firstly find that PDP and each manufacturing parameter have an almost linear correlation as shown in fig.13. This linear correlation resulted in the...
FIGURE 12. Transistor structure optimized by our neural approach with 5% limiter.

TABLE 4. Parameters listed in descending order of sensitivity (dy): values with ‘-’ sign are negatively correlated with PDP.

| Parameter | PDP sensitivity (dy) |
|-----------|----------------------|
| $L_g$     | 6.8710 × 10^{-17}   |
| $N_{sd}(s)$ | 3.7977 × 10^{-17}   |
| $T_{ox}$  | -2.9362 × 10^{-17}  |
| $L_{sdj}(s)$ | 2.1586 × 10^{-17}  |
| $T_{hh}$  | -1.7121 × 10^{-17}  |
| $T_{sp}(s)$ | 1.7022 × 10^{-17}  |
| $N_{nd}(d)$ | -1.3028 × 10^{-17} |
| $N_{halo}(s)$ | -1.2078 × 10^{-17} |
| $L_{sdj}(d)$ | -1.0712 × 10^{-17} |
| $N_{halo}(d)$ | 1.0056 × 10^{-17}  |

Secondly, we find that source/drain parameter pairs have the opposite directional correlations with PDP, but with very similar sensitivities. Highly-doped source region by reducing parasitic resistance and lightly-doped drain region by decreasing gate-induced drain leakage and improving short channel effects (SCEs) are preferred whenever possible [28]–[30]. For example, $L_{sdj}(d)$ and $L_{sdj}(s)$ are negatively and positively correlated with PDP respectively and these two correlations have similar slopes in fig.13. This supports to use asymmetric structure for a better PDP because the positive and negative PDP increments of the same source/drain parameter pair compensate each other.

Thirdly, PDP is insensitive to some manufacturing parameters. For example, PDP is almost constant (dy < 3.066 × 10^{-18}, less than 1% of the optimal predicted PDP 0.3066 × 10^{-15} shown in table.2) for varying $L_{cond}(s/d)$, $L_{halo}(s)$, $N_{ch}$, and $T_{g}$ in fig.13. This means that not all parameters but only some significant parameters with high slope (high sensitivity) are selectively necessary for achieving an optimal PDP [31], [32]. Considering only significant parameters in the transistor optimization process is beneficial because finding an optimum from a lower dimensional search space is faster and easier. Table 4 shows the sorted list of parameters in descending order in regards to their PDP sensitivities (dy) where $L_{g}$, $N_{sd}(s)$, $T_{ox}$, and $L_{sdj}(d)$ are significant factors for PDP optimization based on their PDP sensitivities. This analysis is consistent with the trend of conventional semiconductor industry [33], [34].

We also performed a similar analysis on the effect of manufacturing parameters in delay and power control by monitoring their sensitivities (dy) as shown in fig.14 and fig.15. We categorized manufacturing parameters into five groups regarding their sensitivities (dy) to delay and power, i.e., both insensitive (0,0), delay-only sensitive ($\pm$,0), power-only sensitive (0,$\pm$), both sensitive ($\pm$, $\pm$), both sensitive but opposite ($\pm$, $\mp$). Here, we assumed that parameters of sensitivity (dy) less than 1% of optimal predicted delay (0.1685 × 10^{-11} in table.2) or power (0.1819 × 10^{-3} in table.2) are insensitive. Table 5 shows the parameter groups in descending order of delay and power sensitivities.

TABLE 5. Parameters listed in descending order of delay and power sensitivities (dy): parameters (0,0) are insensitive to both delay and power, Parameters ($\pm$,0) or (0,$\pm$) are sensitive to only delay or power respectively. Parameters ($\pm$, $\pm$) or ($\pm$, $\mp$) are sensitive to both delay and power in the same direction or opposite directions. Each group of parameters are sorted in a descending order in regards to both delay and power sensitivities.

| Parameters (0,0) | Delay sensitivity (dy) | Power sensitivity (dy) |
|-----------------|------------------------|------------------------|
| $L_{cond}(d)$   | 1.9306 × 10^{-14}     | -1.1441 × 10^{-6}      |
| $L_{halo}(d)$   | 1.1255 × 10^{-14}     | 8.5549 × 10^{-7}       |
| $L_{cond}(s)$   | -4.4003 × 10^{-15}    | 1.0203 × 10^{-6}       |
| $T_{g}$         | -1.5111 × 10^{-16}    | 1.5210 × 10^{-6}       |

| Parameters ($\pm$,0) | Delay sensitivity (dy) | Power sensitivity (dy) |
|---------------------|------------------------|------------------------|
| $L_{sdj}(d)$        | -7.5297 × 10^{-14}    | 1.6971 × 10^{-6}       |
| $N_{nd}(d)$         | -7.4103 × 10^{-14}    | 2.5266 × 10^{-7}       |
| $N_{halo}(d)$       | 4.0689 × 10^{-14}     | 1.5459 × 10^{-6}       |
| $L_{sp}(s)$         | 3.5057 × 10^{-14}     | -7.1222 × 10^{-7}      |
| $T_{s/d}(d)$        | 3.2589 × 10^{-14}     | -1.2461 × 10^{-6}      |

| Parameters ($\pm$, $\pm$) | Delay sensitivity (dy) | Power sensitivity (dy) |
|---------------------------|------------------------|------------------------|
| $N_{sd}(s)$               | -4.6829 × 10^{-13}    | 7.1839 × 10^{-6}       |
| $L_{halo}(s)$             | 3.0951 × 10^{-13}     | -4.2408 × 10^{-5}      |
| $L_{rdj}(s)$              | -2.7648 × 10^{-13}    | 5.0168 × 10^{-5}       |
| $L_{sp}(s)$               | 2.2607 × 10^{-13}     | -3.9581 × 10^{-5}      |
| $N_{ch}$                  | 1.8004 × 10^{-13}     | -2.6630 × 10^{-5}      |
| $T_{s/d}(s)$              | 1.1989 × 10^{-13}     | -1.8421 × 10^{-5}      |
| $L_{halo}(d)$             | 7.0730 × 10^{-14}     | -8.4648 × 10^{-6}      |
| $T_{hh}$                  | 5.7565 × 10^{-14}     | -1.4148 × 10^{-5}      |
FIGURE 13. PDP transition for varying manufacturing parameters around the optimal value of 5% limiter and asymmetric source/drain structure.

FIGURE 14. Delay transition for varying manufacturing parameter around the optimal value of 5% limiter and asymmetric source/drain structure.
of sensitivities \(dy\) to delay and power. As we can expect from table 4, \(L_{\text{cond}(d)}\), \(L_{\text{haloj}(d)}\), and \(T_h\) of very low PDP sensitivity are in the both insensitive \((0,0)\) group and have no effect in optimization process. \(N_{\text{ch}}\) and \(L_{\text{haloj}(s)}\) of low PDP sensitivity are categorized as the group of both sensitive but opposite \((\pm, \mp)\) and, therefore, these parameters are useful in controlling delay and power of transistor without changing PDP. By using the parameters in delay-only \((\pm,0)\) sensitive group like \(L_{\text{sd}(d)}\) or power-only \((0,\pm)\) sensitive group like \(T_{\text{ox}}\), we can effectively control delay or power of transistor independently.

Based on this correlation and sensitivity analysis, we find a possibility to further reduce PDP. To this end, we enlarged the range of limiter from 5% of reference manufacturing parameter values to full ranges between Min and Max in table 1 and performed the same automatic parameter optimization process mentioned in sec.III-C. Table 6 shows optimized values of manufacturing parameters with the full range limiter. The resulted parameters positively correlated with PDP as shown in table 4, i.e., \(L_g\), \(N_{\text{haloj}(d)}\), and \(N_{\text{sd}(s)}\), were decreased and parameters negatively correlated with PDP shown in table 4, i.e., \(N_{\text{ch}}\), \(N_{\text{haloj}(s)}\), \(N_{\text{sd}(d)}\), \(T_{\text{th}}\), and \(T_{\text{ox}}\), were increased to the direction of reducing PDP. The other parameters were not changed because they were already in full range with 5% limiter. This full range optimization resulted in about 29% lower PDP \(0.2661 \times 10^{-15}\) than the TCAD simulated value \(0.3749 \times 10^{-15}\) as shown in table 7.

### Table 6. The optimized manufacturing parameters from 40 repeated optimization processes with the full range limiter (Avg: averaged value, Std: standard deviation, Std / Avg: relative deviation): all values were measured from 10 trials of 40 repeated optimization processes.

| Parameter          | Avg     | Std        | Std / Avg  |
|--------------------|---------|------------|------------|
| \(L_{\text{cond}(d)}\) | 42      | 4.5336 \times 10^{-7} | 1.0801 \times 10^{-5} |
| \(L_{\text{cond}(s)}\) | 38      | 8.1926 \times 10^{-7} | 2.1559 \times 10^{-5} |
| \(L_g\)            | 25.6    | 1.3720 \times 10^{-8} | 1.4442 \times 10^{-6} |
| \(L_{\text{haloj}(d)}\) | 9.5     | 4.2806 \times 10^{-8} | 4.0768 \times 10^{-6} |
| \(L_{\text{haloj}(s)}\) | 10.5    | 4.2387 \times 10^{-8} | 1.6147 \times 10^{-6} |
| \(L_{\text{sd}(d)}\) | 26.2    | 1.5486 \times 10^{-12} | 6.5204 \times 10^{-11} |
| \(L_{\text{sd}(s)}\) | 23.8    | 6.6142 \times 10^{-8} | 3.4812 \times 10^{-6} |
| \(L_{\text{sh}(d)}\) | 19      | 4.0887 \times 10^{-11} | 1.9470 \times 10^{-9} |
| \(L_{\text{sh}(s)}\) | 21      | 2.4293 \times 10^{-13} | 1.9435 \times 10^{-5} |
| \(N_{\text{ch}}\) | 1.2500 \times 10^{18} | 2.8333 \times 10^{7} | 7.5556 \times 10^{-12} |
| \(N_{\text{haloj}(d)}\) | 3.7500 \times 10^{18} | 1.1724 \times 10^{13} | 1.8759 \times 10^{-9} |
| \(N_{\text{haloj}(s)}\) | 6.2500 \times 10^{18} | 2.5697 \times 10^{6} | 2.0558 \times 10^{-14} |
| \(N_{\text{sd}(d)}\) | 1.2500 \times 10^{20} | 1.9691 \times 10^{6} | 1.6255 \times 10^{-15} |
| \(N_{\text{sd}(s)}\) | 0.7500 \times 10^{20} | 8.0618 \times 10^{-7} | 1.6972 \times 10^{-5} |
| \(T_g\)            | 47.5    | 4.5714 \times 10^{-10} | 1.9876 \times 10^{-16} |
| \(T_{\text{th}}\)   | 23      | 1.6864 \times 10^{-8} | 3.0326 \times 10^{-6} |
| \(T_{\text{sh}(d)}\) | 9.5     | 1.6864 \times 10^{-8} | 1.6061 \times 10^{-6} |
| \(T_{\text{sh}(s)}\) | 10.5    | 1.6864 \times 10^{-8} | 1.6061 \times 10^{-6} |

### E. COMPARISON BETWEEN HUMAN EXPERT AND OUR NEURAL APPROACH

For a fair comparison of optimization performances between our neural optimization method and human expert, we set the same search ranges of manufacturing parameters, i.e., 5% limiter (sec.III-C) and full range limiter (from Min to Max of
TABLE 7. TCAD simulated values of PDP, delay, and power for the optimized manufacturing parameters: ‘Neural’ represents the optimization with our neural models and asymmetric source/drain structure, ‘Neural sym’ with our neural models and symmetric source/drain structure, ‘Human expert’ with human optimization and symmetric source/drain structure, and ‘Reference structure’ shows the outputs of TCAD simulation with the reference parameter values [8] in table 1.

| Optimization method | PDP $\times 10^{-15}$ | Delay $\times 10^{-14}$ | Power $\times 10^{-3}$ |
|---------------------|------------------------|-------------------------|-----------------------|
| Neural (5% limiter) | 0.3120                 | 0.1099                  | 0.1836                |
| Neural sym (5% limiter) | 0.3535               | 0.1409                  | 0.2509                |
| Human expert (5% limiter) | 0.3487               | 0.1435                  | 0.2430                |
| Neural (full range)  | 0.2661                 | 0.2272                  | 0.1171                |
| Neural sym (full range) | 0.3014               | 0.1490                  | 0.2023                |
| Human expert (full range) | 0.2865               | 0.1445                  | 0.2082                |
| Reference structure [8] | 0.3749               | 0.1516                  | 0.2473                |

In case of our neural approach, we performed additional automatic parameter optimization with our trained models, previously mentioned in sec.III-B as described in sec.III-C but with source/drain symmetric constraint, i.e., using a single parameter for each pair of source/drain parameters which has a symbol with ($s/d$) in table 1.

In case of human expert, an experienced human expert optimized the PDP of the conventional MOSFET as described here: to minimize the PDP, gate capacitance $C_{gg}$ should be minimized according to the PDP, product of delay and power. Gate length was first tuned to reduce the $C_{gg}$. Then, equivalent oxide thickness (EOT), directly related to $T_{ox}$ and $T_{hk}$, was controlled. In this process, it was revealed that the EOT does not affect the SCEs of the device in this work, so a large EOT was set to reduce the $C_{gg}$. Both $T_{g}$ and $T_{sd}$ were decreased to minimum in order to reduce the fringing capacitance between gate stack and S/D regions. $L_{con}$ was also decreased to minimum in order to reduce the junction capacitance. Doping-related parameters, $N_{ch}$, $N_{halo}$, and $N_{sd}$, were decreased to reduce the overlap and intrinsic channel capacitances. Finally, $L_{sp}$ was set to the middle value due to the trade-off between overlap and fringing capacitances.

1) COMPARISON IN FIGURE-OF-MERITS: PDP, DELAY, AND POWER

Table 7 and fig.16 present and compare figure-of-merits of transistors optimized by our neural approaches and human expert. In case of 5% limiter, both human expert and our neural method (Neural sym) achieved about 7% less PDP and delay and similar power compared to those resulted from the reference structure. In case of full range limiter, both methods achieved about 20% less PDP and power and slightly less delay than those resulted from the reference structure.

When comparing results from human expert and our method (Neural sym), both showed similar delays, and powers; while PDPs were also similar, our method resulted in a slightly higher PDP compared to that resulted by human expert. Our method has 1 to 2% error in modeling (fig.9 in sec.III-B) and optimization (fig.10 in sec.III-C). And, due to the linear correlation of low complexity (fig.13, fig.14, and fig.15), human expert easily found the near best manufacturing parameters. This resulted in a lower PDP in Neural sym but the difference was not significant, coming to less than 1.5%.

We presented optimized manufacturing parameters by our neural approach under the constraint of symmetric source/drain (Neural sym).
TABLE 9. Optimized manufacturing parameters by human expert with conventional MOSFET of symmetric source/drain.

| Parameter          | Value (5% limiter) | Value (full range) |
|--------------------|--------------------|--------------------|
| $L_{con(s/d)}$    | 38                 | 40                 |
| $L_g$              | 28.5               | 25.5               |
| $L_{halo(s/d)}$    | 9.5                | 9.5                |
| $L_{dy(s/d)}$      | 23.75              | 23.8               |
| $L_{epi(s/d)}$     | 19                 | 20                 |
| $N_{ch}$           | $0.9500 \times 10^{18}$ | $0.7500 \times 10^{18}$ |
| $N_{halo(s/d)}$    | $4.7500 \times 10^{18}$ | $3.7500 \times 10^{18}$ |
| $N_{sd(s/d)}$      | $0.9500 \times 10^{20}$ | $0.7500 \times 10^{20}$ |
| $T_g$              | 47.5               | 47.5               |
| $T_{hk}$           | 2.1                | 2.3                |
| $T_{ox}$           | 0.735              | 0.805              |
| $T_{s/d(s/d)}$     | 9.5                | 9.5                |

TABLE 10. Parameters listed in descending order of sensitivity (dy) of symmetric source/drain structure: values with ‘-’ sign are negatively correlated with PDP. (See appendix.A for the corresponding transitions of figure-of-merits).

| Parameter          | PDP sensitivity (dy) | Parameter          | PDP sensitivity (dy) |
|--------------------|----------------------|--------------------|----------------------|
| $L_g$              | $9.6238 \times 10^{-17}$ | $L_{con(s/d)}$  | $-4.8686 \times 10^{-18}$ |
| $T_{ox}$           | $-3.7973 \times 10^{-17}$ | $L_{epi(s/d)}$  | $4.7325 \times 10^{-18}$ |
| $T_{hk}$           | $-2.6635 \times 10^{-17}$ | $T_g$           | $-4.5583 \times 10^{-18}$ |
| $N_{ch}$           | $8.5829 \times 10^{-18}$ | $L_{halo(s/d)}$ | $4.0593 \times 10^{-18}$ |
| $N_{sd(s/d)}$      | $5.4069 \times 10^{-18}$ | $N_{halo(s/d)}$ | $2.9812 \times 10^{-18}$ |
| $L_{adj(s/d)}$     | $-5.3249 \times 10^{-18}$ | $T_{s/d(s/d)}$  | $1.4272 \times 10^{-18}$ |

analysis in symmetric source/drain structure (table 10). The other parameters have the same or similar values in both methods. These results demonstrate that our neural approach achieved a comparable optimization result to the equation-based minimization of human expert without knowing any equation through the entire process. In the case of full range limiter, as the search range increased from 5% of reference parameter values to the full range in table 1, the parameters positively correlated with PDP (table 10), i.e., $N_{ch}$, $N_{halo(s/d)}$, and $N_{sd(s/d)}$, were decreased and the parameters negatively correlated with PDP (table 10), i.e., $T_{hk}$ and $T_{ox}$, were increased in order to get minimal PDP in both methods (table 8 and table 9).

2) COMPARISON OF SYMMETRIC/ASYMMETRIC STRUCTURES

We find that our neural method with asymmetric structure (Neural) optimized PDP much less than with symmetric structure (Neural sym) as shown in table 7 and fig.16. As mentioned in sec.III-C, two parameters in a source/drain pair are oppositely correlated to PDP and this resulted in a higher PDP in symmetric structure. In addition, there is higher chances to find better parameters to obtain a lower PDP in asymmetric structure since asymmetric structure has 19 degree-of-freedom (DOF), which has seven more variables than symmetric structure with 12 DOF.

3) COMPARISON IN OPTIMIZATION SPEED

Table 11 shows the elapsed time to get optimal manuscript parameters with our neural method and human expert. Our neural method consumed 80 seconds for training two neural models described in sec.III-B and another 80 seconds for automatically optimizing manufacturing parameters described in sec.III-C. Once the models are trained, then our method only required 80 seconds for additional optimization. In contrast, human expert consumed 3 to 4 hours to obtain optimal manufacturing parameters because of a lot of trial & error tuning with TCAD simulations. Although these speeds are not the most optimized for both methods, we can roughly conclude that our neural method is about 100 times faster (2 to 3 minutes) than that of human expert (180 to 240 minutes).

Our neural approach requires a thousand of training samples generated by TCAD simulations. However, this data generation is off-line process which can be done by random sampling without delicate grid point control and in parallel utilizing multi-cores within several sampling time. Because of this, data acquisition time is not usually significantly considered in machine-learning-related works. Similarly, there are several papers about neural approach for the optimization of devices and circuits which do not consider the data acquisition time either [16], [35], [36]. Even when considering data acquisition time, our neural method requires only several TCAD simulation time for generating whole training samples through parallel computing while human expert needs lots of sequential simulation time because of a bundle of sequential trial and error tuning processes. In our experiment, TCAD simulation time for a data sample required 1 minute and we used 3 parallel simulations only because of license limitation. Although this resulted in 333 minutes (about 5.5 hours) to obtain 1000 training samples, if we use more number of parallel simulations, data acquisition time can be reduced down to 1 min (when 1000 parallel simulations are used), a single data sample time.

IV. CONCLUSION

In this paper, we proposed a new framework to automatically optimize manufacturing parameters of transistor by using neural networks. As the first step, our method trained two neural networks for delay and power models of a transistor utilizing a training data set from TCAD simulation. Then, an optimal set of manufacturing parameters that minimizes PDP was found automatically by using the neural network models and gradient descent method in desired ranges of parameters.

We experimentally proved the superiority and efficiency of our neural approach compared to the previous Intel 32 nm node [8] and human expert’s MOSFET optimization in symmetric source/drain structure. Figure-of-merits of our method were significantly better than the TCAD simulated values with the reference structure [8], i.e., about 7% less PDP with 5% limiter and about 20% less PDP with full range limiter. Additionally, Figure-of-merits of our method were comparable to human expert’s optimization resulted via time-consuming trial & error with TCAD simulations, while our method only took a couple of minutes to do the same optimization.
TABLE 11. Timing comparison of our neural approach and human expert: timing of our method (Neural or Neural sym) was measured with MATLAB script (not optimized) and timing of human expert was roughly measured in hour. N represents the number of parallel processes used in training data generation.

| Optimization method | Time consumption |
|---------------------|------------------|
| Neural or Neural sym| \( \frac{N}{1 \text{ min (data acquisition)}} + 80 \text{ s (model training)}} + 80 \text{ s (optimizing)} = \frac{160}{N} + 2.7 \text{ min} |
| Human expert        | 3 to 4 hours     |

Moreover, it was possible to analyze the effect of each manufacturing parameter on figure-of-merits when our neural models were utilized. Without a huge amount of TCAD simulations, our neural models obtained sensitivities of parameters to PDP, delay, and power by predicting these figure-of-merits for varying parameters. Then, we categorized and gave certain weights to manufacturing parameters regarding their sensitivities to PDP, delay, and power. For example, \( L_g \), \( T_{ox} \), \( T_{hk} \) are highly correlated to PDP but \( L_{cond}(s/d) \) and \( T_g \) are uncorrelated to any of PDP, delay, and power.

Our neural approach is not limited to a specific device and matured technology but can be expanded to general devices and new technology as long as a number of input/output

FIGURE 17. PDP transition for varying manufacturing parameter around the optimal value of 5% limiter and symmetric source/drain structure.

FIGURE 18. Delay transition for varying manufacturing parameter around the optimal value of 5% limiter and symmetric source/drain structure.
data are given because machine learning itself does not use any device physics considering device structure and scale.

Even though we did not perform real device experiments but focused on simulated results as done in the machine-learning-based circuit designing method [16] which includes only simulated results without implementing real circuit device, based on our experience, the simulation results are rather correct and well matched to the real devices.

There would be concerning about gathering training data from real factory. In real factory, for variation in unit manufacturing process, they cannot examine all wafer but some dies as representative samples. Our work is focusing on optimizing variation in unit manufacturing process with the
varying range from min to max in Table 1 and, in the same manner, we used 1000 data samples for machine learning which can also be obtained in real factory without heavy expense. For output data, it would be burdensome if we measure full $I_{ds}$-$V_{gs}$ and $C_{gg}$-$V_{gs}$ data through gate voltage sweep, but several $I_{ds}$ and $C_{gg}$ values at some specific gate voltages.
can be obtained in real factory with no heavy expense. For input data, process parameters can be obtained from the devices at the scribe line. Randomized structural parameters in Table 1 are the outcomes of the process parameters such as dose, temperature, and gas flow, but the accuracy of machine learning is not affected by the type of inputs. If there are some missing data, we can adopt generative adversarial network (GAN) as a realistic data generator to complete them.

Our experiments demonstrated that PDP of transistor can be much lower to about 29% less than that of the reference structure, if asymmetric source/drain structure is adopted. In addition, PDP can be further improved with even wider limiter beyond Min and Max shown in table 1 because linear transitions of figure-of-merits for full range limiter exhibit monotonic functions (see figures in appendix B and C). Since our neural approach is not limited to a specific figure-of-merit or scale of device, we can apply our method to a smaller device for better figure-of-merits and to control each figure-of-merit, such as lowering down the relatively high delay in asymmetric structure in table 7 by defining an appropriate objective function alternative to eq.3. There is a need to further analyze asymmetric structure and wider limiter for real application because they are not compatible with the previous 32 nm node with symmetric structure, and subsequently, there...
is an additional fabrication cost for asymmetric structure. This is remained as our future work.

**APPENDIX. TRANSITIONS OF FIGURE-OF-MERITS**

**A. SYMMETRIC STRUCTURE AND 5% LIMITER**  
See Fig. 17 to 19 here.

**B. ASYMMETRIC STRUCTURE AND FULL RANGE LIMITER**  
See Fig. 20 to 22 here.

**C. SYMMETRIC STRUCTURE AND FULL RANGE LIMITER**  
See Fig. 23 to 25 here.

**REFERENCES**

[1] A. Khakifirooz, O. M. Nayfeh, and D. Antoniadis, “A simple semiempirical short-channel MOSFET current–voltage model continuous across all regions of operation and employing only physical parameters,” *IEEE Trans. Electron Devices*, vol. 56, no. 8, pp. 1674–1680, Aug. 2009, doi: 10.1109/T-ED.2009.2024022.

[2] BSM Models. [Online]. Available: http://bsim.berkeley.edu/models/

[3] Predictive Technology Model. [Online]. Available: http://ptm.asu.edu/

[4] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford Univ. Press, 1995.

[5] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[6] A. Ceyhan, J. Quijas, S. Jain, H.-Y. Liu, W. E. Gifford, and S. Chakravarty, “A physically based mobility model for numerical simulation of nonplanar devices,” in *IEEE Trans. Computer-Aided Design Integr. Circuits Syst.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[7] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[8] A. Ceyhan, J. Quijas, S. Jain, H.-Y. Liu, W. E. Gifford, and S. Chakravarty, “A physically based mobility model for numerical simulation of nonplanar devices,” in *IEEE Trans. Computer-Aided Design Integr. Circuits Syst.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[9] K.-I. Funahashi, “On the approximate realization of continuous mappings by neural networks,” *Neural Netw.*, vol. 2, no. 3, pp. 183–192, Jan. 1989.

[10] P. Werbos, “Beyond regression: ‘New tools for prediction and analysis in the behavioral sciences,”’ Ph.D. dissertation, Dept. Appl. Math., Harvard Univ., Cambridge, MA, USA, 1975.

[11] C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi, “A physically based mobility model for numerical simulation of nonplanar devices,” in *IEEE Trans. Computer-Aided Design Integr. Circuits Syst.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.
HYEOK YUN (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering and the Ph.D. degree in creative IT engineering from the Pohang University of Science and Technology (POSTECH), South Korea, in 2012 and 2016, respectively. He was a Postdoctoral Research Fellow with POSTECH from 2016 to 2018. Since 2019, he has been a Research Assistant Professor of electrical engineering with POSTECH. His research interests include characterization and simulation of advanced nanoscale devices (fin, gate-all-around, tunneling, and nanosheet FETs) and applications (chemical sensor and solar cell).

JUN-SIK YOON (Member, IEEE) received the B.S. degree in electrical engineering and the Ph.D. degree in creative IT engineering from the Pohang University of Science and Technology (POSTECH), South Korea, in 2012 and 2016, respectively. He was a Postdoctoral Research Fellow with POSTECH from 2016 to 2018. Since 2019, he has been a Research Assistant Professor of electrical engineering with POSTECH. His research interests include characterization and simulation of advanced nanoscale devices (fin, gate-all-around, tunneling, and nanosheet FETs) and applications (chemical sensor and solar cell).

ROCK-HYUN BAEK (Member, IEEE) received the B.S. degree in electrical engineering from Korea University, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2006 and 2011, respectively. He was a Postdoctoral Researcher and a Technical Engineer with SEMATECH, Albany, NY, USA, from 2011 to 2015. He was a Senior Device Engineer with the SAMSUNG Research and Development Center (Pathfinding TEAM), South Korea, from 2015 to 2017. Since 2017, he has been an Assistant Professor of electrical engineering with POSTECH. His research interests include technology benchmark by characterization, simulation, and modeling of advanced devices and materials (fin, gate-all-around, nanosheet FETs, 3D-NAND, 3DIC, SiGe, Ge, III-V, and so on).

* * *

HYUN-CHUL CHOI received the B.S. degree in electronic and electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2002, and the M.S. and Ph.D. degrees in electronic and electrical engineering from the Pohang University of Science and Technology (POSTECH), South Korea, in 2005 and 2011, respectively. From 2011 to 2014, he was a Research Scientist with the Multimedia Development Laboratory of Daum Communications. He has been an Assistant Professor with the Department of Electronic Engineering, Yeungnam University, South Korea, since 2014, where he has also been an Associate Professor since 2020. His research interests include multimedia retrieval, computational photography, 3D photogrammetry, object tracking and recognition based on machine learning, and neural networks.