Received 9 November 2023; revised 14 December 2023; accepted 20 December 2023. Date of publication 22 December 2023;
date of current version 29 January 2024. The review of this article was arranged by Editor J. Wang.

Digital Object Identifier 10.1109/JEDS.2023.3346380

Enhancement and Expansion of the Neural Network-Based Compact Model Using a Binning Method

JINYOUNG CHOI [1] (Graduate Student Member, IEEE), HYUNJOON JEONG [2] (Graduate Student Member, IEEE), SANGMIN WOO [2], HYUNGMIN CHO [1] (Student Member, IEEE), YOHAN KIM [3] (Member, IEEE), JEONG-TAEK KONG [4], AND SOYOUNG KIM [4] (Senior Member, IEEE)

1 Department of Semiconductor and Display Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
2 Department of Electrical and Computer Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
3 Computational Science and Engineering Team, Innovation Center, Samsung Electronics, Suwon 16677, Republic of Korea
4 Department of Semiconductor Systems Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea

CORRESPONDING AUTHORS: S. KIM AND J. KONG (e-mail: ksyoung@skku.edu; jtkong@skku.edu)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government (MSIT, Software Systems for AI Semiconductor Design) under Grant 2021-0-00754, and in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant 2020R1A5A1019649.

ABSTRACT The artificial neural network (ANN)-based compact model has significant advantages over physics-based standard compact models such as BSIM-CMG because it can achieve higher accuracy over a wide range of geometric parameters. This makes it particularly suitable for design space exploration and optimization. However, the ANN-based compact model using only one set of model parameters (global-ANN) requires larger model sizes to achieve wider coverage and higher accuracy in order to capture the unpredictable nonlinearities of emerging devices. This results in reduced simulation speed and a trade-off between simulation accuracy, model coverage, and simulation speed makes it difficult to utilize ANN-based compact models in a variety of ways. To solve this problem, we propose the first ANN-based compact modeling flow using a binning method (binning-ANN) and we address the training requirements and data sparsity issues that may occur due to the binning method in ANNs. In addition, we develop a bin size optimization guideline for the binning-ANN. As a result, the binning-ANN not only has higher accuracy, but also much better expandability than existing methods.

INDEX TERMS Artificial neural network (ANN), machine learning (ML), device modeling, compact model, binning, emerging device, SPICE.

I. INTRODUCTION

Fast and accurate models of next-generation devices are very important for circuit simulation and design optimization [1]. Industry-standard compact models such as the Berkeley short-channel IGFET model (BSIM) have used physics-based equations to describe electrical characteristics of devices [2], [3]. However, developing a physics-based compact model is very time-consuming because of the many secondary effects such as short-channel effects and quantum effects in emerging devices. In addition, since the standard compact model is mainly an analytical model, it can only predict the electrical characteristics over a very narrow range [4]. A compact model should have wide model coverage and allow the use of a wide range of design parameters for simulation. An artificial neural network (ANN) model is used to efficiently predict and analyze nonlinear electrical characteristics according to the process parameters of the device [5]. Due to these advantages, the ANN-based compact model is suitable for iterative design optimizations and parametric analysis because it effectively predicts nonlinear electrical characteristics for geometric, physical, material, and bias parameters [6].

For iterative design space explorations and optimizations, it is very important to design a compact model with high accuracy, wide model coverage, and fast simulation speed. However, we found a critical issue in previous studies in which increasing the number of neurons and layers of the ANN model increased the amount of computation and slowed down the simulation speed [7], [8]. In general, to predict the nonlinear electrical characteristics of many devices with high accuracy, hyperparameters such as neurons and layers of the ANN model must be increased, which increases the simulation time. This is the first trade-off between simulation speed and accuracy of compact models. Another consideration is that the ANN model is only accurate in the range for which it was trained. To use a compact model for new devices according to changes in node technology, the new devices must be added to the training data. If the electrical characteristics of the new devices have similar tendencies to the existing trained devices, transfer learning can be used to achieve high accuracy with a small set of training data and training times [9], [10]. However, when new training data are added, the accuracy of the ANN model decreases because the regression problem becomes more complex. This is the second trade-off between accuracy and coverage of compact models. For the first time, we present the idea of using multiple sets of model parameters in an ANN-based compact model as a method to address the abovementioned two issues.

For several years, ANN-based compact models called 'global models' used only one set of model parameters for all devices. Since it is very difficult to capture all nonlinear characteristics in emerging devices with one set of model parameters in the industry standard compact model, multiple sets of model parameters are used to separate intervals with similar characteristics to improve accuracy [11], [12]. This is called the binning method. However, parameter extraction of the standard compact model takes a lot of time, making the binning method inefficient [8], [11], [13], [14]. Unlike the standard compact model, adopting a binning method to ANN-based compact models can be efficient because the required time cost and modeling efforts are very small [8]. There are many studies on various training and preprocessing methods to improve the efficiency of neural network models. However, there is a lack of research on an efficient ANN-based compact model framework that simultaneously considers the ANN and the Verilog-A language. As shown in Table 1, we introduce a technique to strengthen the model by applying the binning method to the ANN-based compact model. The contributions of this study are as follows.

1) We applied the binning method to data-driven ANN models for the first time to maximize learning efficiency.
2) We proposed a novel training method and a bin size optimization flow using transfer learning to address the increased training requirements and data sparsity issues that may arise from the binning method.

TABLE 1. Type of compact models.

Compact Model Type	Standard Compact Model		ANN-based Compact Model	
Based Equation	Device Physics		Neural Network	
Model Parameter Extraction Methodology	Global	Binning	Global	**Binning**
Model Development Time	High	High	Low	**Low**
Accuracy	Low	High	Medium	**High**
Modeling Efforts	High	Very High	Very Low	**Low**
Model Coverage	Narrow	Unrestricted	Wide	**Unrestricted**

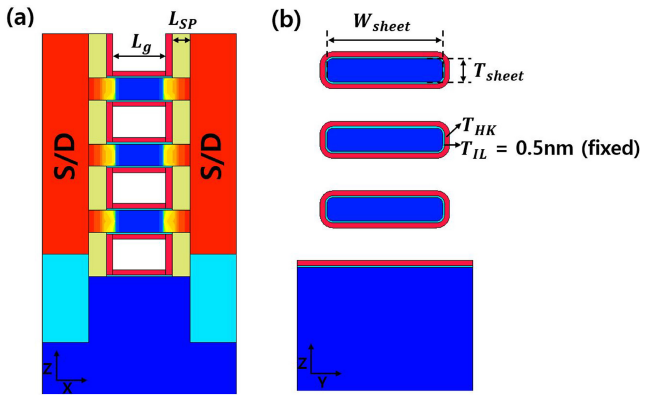

FIGURE 1. (a) The NSFET y-axis section, (b) The NSFET x-axis section.

3) Using Verilog-A, we constructed an ANN-based compact model that uses multiple sets of model parameters to address the inherent trade-offs between simulation speed, ANN accuracy, and available device coverage.

In Section II, we propose a binning method to improve the accuracy of ANNs in both trained and testing devices. In Section III, we introduce the model integration method using Verilog-A and show the unrestricted and fast model expansion method using the binning method and transfer learning. In Section IV, the performance of the proposed binning method is evaluated for digital and analog circuits. Section V summarizes the study and discusses future work.

II. BINNING METHOD IN ANNS

A. TECHNOLOGY COMPUTER-AIDED DESIGN (TCAD) SIMULATION

As shown in Fig. 1, the device used for the ANN-based compact model is a three-stacked Nanosheet FET (NSFET) as a next-generation device. Simulations are conducted using Synopsys Sentaurus TCAD. The NSFET simulation conditions are the same as those of our previous studies [7], [8]. When $V_{ds} = 0.65$V, 53 I-V points were uniformly extracted from the range of $V_{gs} = 0 \sim 0.65$V. The T_{IL} (i.e., the SiO_2 ($k = 3.9$) thickness) was fixed at 0.5 nm and the equivalent oxide thickness (EOT) was swept while changing the T_{HK} (i.e., the HfO_2 ($k = 22$) thickness). The inner spacer length (L_{sp}), gate length (L_g), sheet width (W_{sheet}), and sheet thickness (T_{sheet}) were swept. When extracting data, sampling methods considering the distribution of data such as Latin hypercube sampling (LHS) and Sobol sequence are more effective than uniform sampling [15]. In the range of

TABLE 2. NSFET dataset.

Type	Symbol	Parameters	Range	Distribution	Spearman
Geometric	L_g	Gate length	11 ∼ 16 [nm]	Sobol	-0.083
Geometric	L_{sp}	Spacer thickness	3 ∼ 5 [nm]	Sobol	-0.0439
Geometric	T_{sheet}	Sheet thickness	4 ∼ 6 [nm]	Sobol	0.137
Geometric	W_{sheet}	Sheet width	20 ∼ 50 [nm]	Sobol	0.0817
Geometric	EOT	Equivalent oxide thickness	0.59 ∼ 0.77 [nm]	Sobol	0.00635
Bias	V_{gs}	Gate-source voltage	0 ∼ 0.65 [V]	Uniform	N/A
Bias	V_{ds}	Drain-source voltage	0.65 [V]	Fixed value	N/A

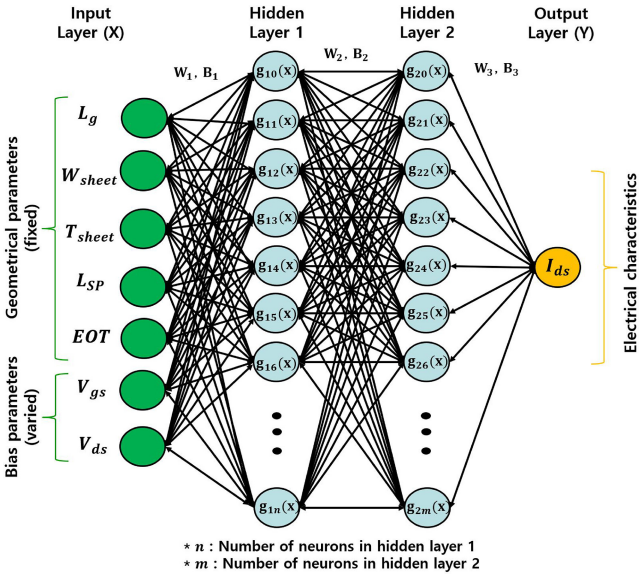

FIGURE 2. The ANN model architecture.

Table 2, 1,024 devices were generated using TCAD's Sobol sequence [16]. The total number of device data points is 54,272.

B. ANN MODEL

The algorithm and structure of the ANN model are similar to those of previous studies [7], [8]. Early stopping was also used to prevent overfitting. We used an NVIDIA Titan Xp graphics card to accelerate the training of the ANN model. The ANN model uses the hyperbolic tangent as an activation function. As shown in Fig. 2, the model that predicts the drain current according to geometric parameters and bias parameters is trained. Since the ANN model should have the ability to fit electrical characteristics such as I_{ds}, C_{gg}, C_{gd}, and C_{gs} while setting the size as small as possible for the speed of SPICE simulation. The ANN model has two hidden layers with 10 neurons to achieve a target accuracy greater than 95%. As shown in Equation (1), the ANN model uses a fixed MinMaxScaler because the formula can be changed as the range of the dataset changes due to the binning method.

$$X_i = \frac{x_i - min_{fixed}(x)}{max_{fixed}(x) - min_{fixed}(x)} \tag{1}$$

In this study, a mean square error (MSE)-based physics-augmented loss function was used to perform optimized analysis with the smallest error in the target operation region. As shown in Equation (2), the SE calculation of each region

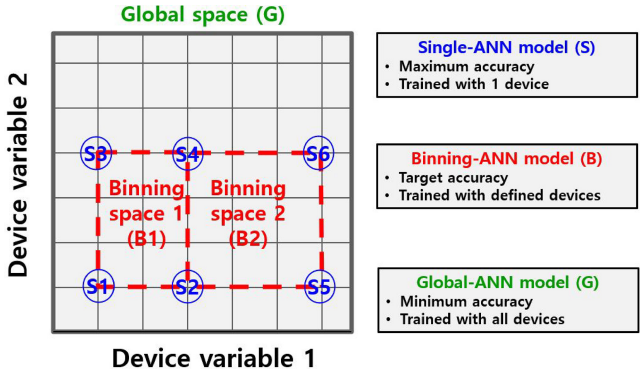

FIGURE 3. Feasible windows of the ANN-based compact model.

was multiplied by α, β, or γ to further reduce the errors in regions with large weights.

$$Loss = \frac{1}{N}\sum_{i=1}^{N_s}\big[\alpha \times SE(I_{off}) + \beta \times SE(I_{triode}) + \gamma \times SE(I_{sat})\big] \tag{2}$$

(SE: square error, α: weight parameter of the off region, β: weight parameter of the triode region, γ: weight parameter of the saturation region) Both the global-ANN and the binning-ANN used the same ANN structure and hyperparameters (e.g., activation function, model size, learning rate) described above.

C. OVERALL ANN-BASED COMPACT MODELING FLOW WITH BINNING

The binning method used for compact models means dividing an entire range of devices into smaller groups or "bins" based on similar electrical characteristics to increase fitting accuracy. The same method can be applied to the ANN-based compact model. Fig. 3 shows the feasible windows of the ANN model that can be designed according to the range of devices used for training. In our previous study [7], we analyzed the advantages and disadvantages of the single-ANN model and the global-ANN model. The single-ANN model refers to an ANN model that uses only one device for learning and has the highest accuracy, but geometric parameter sweep simulation is not possible. On the other hand, the global-ANN model can be used for all devices but has low model accuracy. The binning-ANN model is a hybrid model that combines the advantages of both models. It can have higher accuracy than the global-ANN by grouping devices with similar electrical characteristics. Because the training set of the model is divided into bins that are closer to a single-ANN model, simulation accuracy improves, but the number of ANN models that need to be trained increases, ultimately increasing the modeling time and training data required to create the model.

Fig. 4 describes the proposed ANN-based compact modeling flow using the binning method. The device data is generated by TCAD simulations or real devices. The key to

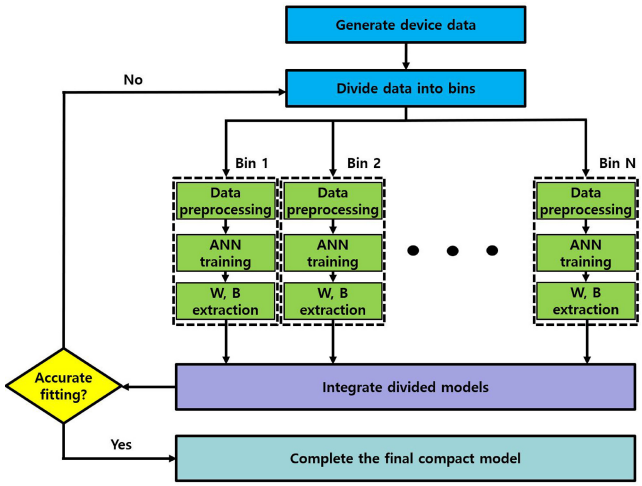

FIGURE 4. The proposed ANN-based compact modeling flow with binning.

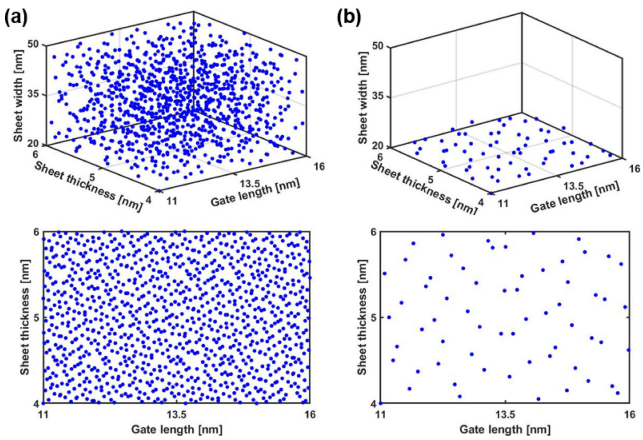

FIGURE 5. Data distribution of (a) 1,024 devices, (b) 64 devices.

improving model accuracy in the ANN-based compact model with the binning flow is training data splitting. By splitting the original training data to generate an easy regression problem, the accuracy of the ANN model can be dramatically improved without increasing the size of the model. The divided ANN models are trained independently. After model training, each model is evaluated with both training devices and testing devices. If the accuracy of the integrated model is higher than the target accuracy, we complete the compact model. If not, the training data are divided into smaller bins, and the process is repeated.

D. GLOBAL TO BINNING FINE-TUNING (GTBF) METHOD

The binning method in ANN can lead to two drawbacks. First, the need to generate an ANN model for each bin leads to increased training requirements as mentioned above. Second, the method can lead to a data sparsity issue. Fig. 5 shows the data distribution of (a) 1,024 device sets and (b) 64 device sets divided into 16 equal bins with the W_{sheet} as the binning parameter. As shown in Fig. 5, increasing the number of bins is equivalent to decreasing

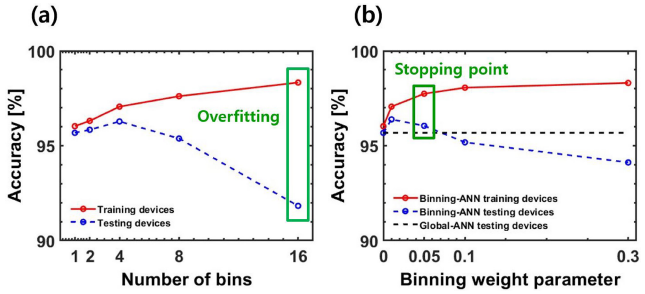

FIGURE 6. (a) The influence of the number of bins on model accuracy, (b) The influence of binning weight parameter (W_B) on model accuracy.

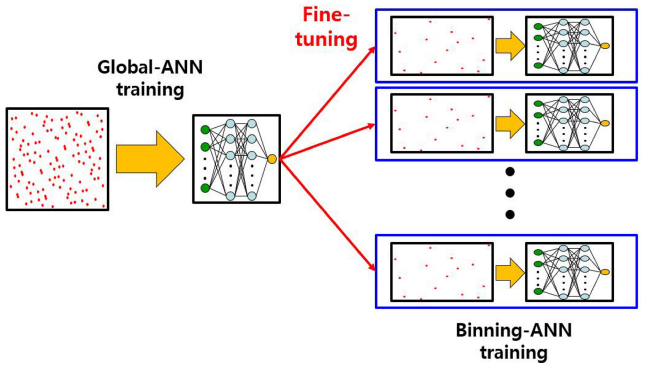

FIGURE 7. Block diagram of the global to binning fine-tuning method (GTBF).

the bin size so that the density of training devices decreases rapidly. Fig. 6 (a) shows the accuracy of the training devices and testing devices according to the number of bins. Since the ANN model is data-driven with multi-dimensional inputs, it becomes more vulnerable to overfitting when the number of training devices decreases due to the binning method. To address these issues, we propose a new training method. In the field of device modeling, there are studies in which transfer learning is used to alleviate the problem of insufficient training data [9], [10]. Fine-tuning is one of the methods of transfer learning and is an algorithm used to make small modifications to suit a new purpose based on the existing ANN. Fig. 7 shows the proposed global to binning fine-tuning method (GTBF) to resolve the problem of overfitting. To design an optimized binning-ANN, we maximize the accuracy of the training devices while maintaining a higher accuracy of the testing devices as compared to the global-ANN. Before training the binning-ANN, the global-ANN with the original training data is first trained. We use the global-ANN as an initial binning-ANN and continue training with a very low learning rate. The GTBF method has three advantages over normal ANN training methods. First, because a global-ANN with a relatively high density of training data is used as the initial model, it can have higher accuracy for the testing devices than normal binning-ANN. Second, fine-tuning can greatly reduce the training requirements using a pre-trained ANN model. Equation (3) is the training requirements for a normal

binning-ANN training method and Equation (4) is the epoch requirement for GTBF. As shown in Equation (5), because the training is rapidly accelerated using transfer learning, $T_{fine-tuning}$ is mostly very small ($W_B < 0.3$) compared to T_{global}.

$$T_{normal} = N_{bin} \times T_{global} \tag{3}$$

$$T_{GTBF} = T_{global} + N_{bin} \times T_{fine-tuning} \tag{4}$$

$$T_{fine-tuning} = W_B \times T_{global} \tag{5}$$

(N_{bin}: number of bins, T_{global}: epoch requirement for global-ANN, W_B: binning weight parameter in Fig. 6 (b))

As the number of bins increases, the difference between T_{normal} and T_{GTBF} increases. The training requirements can become less than 30% of normal binning-ANN. Lastly, and most importantly, overfitting of the ANN model can be prevented by continuously comparing the binning-ANN with the global-ANN during training. As shown in Fig. 6 (b), the GTBF method can be used as a condition to stop learning. It prevents overfitting due to the data sparsity problem by comparing the accuracy of the global-ANN and the binning-ANN on testing devices at specific intervals. As the binning weight parameter increases, the accuracy of the training devices continues to increase, but the accuracy of the testing devices increases and then decreases.

E. DECISION CRITERIA FOR BINS

There are two steps to design the binning-ANN model. The first step is to select the parameters for binning. The inputs of the ANN-based compact model are divided into fixed parameters such as geometric, physical, material, temperature, and process parameters which have fixed values in DC, AC, and transient simulations, and variable parameters such as voltage biases whose values continuously change during the simulation. Continuity and smoothness may not be guaranteed when the model parameters are changed at the bin boundaries. To avoid convergence problems during SPICE simulations, binning must be performed to the parameters that are fixed. Among the geometric parameters, we selected five key parameters of three-stacked nanosheet FETs [7]. Since the main idea of the binning method is to group devices with similar characteristics into a same bin, it is advantageous to apply the binning method using input parameters that have a large dependence on the output. We analyzed this dependence by finding Spearman correlation coefficients. The Spearman correlation coefficient shows dependence between two parameters in the data. As shown in Table 2, we extracted the Spearman correlation coefficients between the preprocessed input parameters and the output. T_{sheet}, L_g, and W_{sheet} were used to create bins.

The second step is to select the proper number of bins and the size of each bin. By dividing the input data into bins, the range of training data that the ANN must fit will be reduced, then the accuracy of the model will increase. Fig. 8 shows a plot of the validation loss according to the number of bins and training epochs. It shows that achievable accuracy

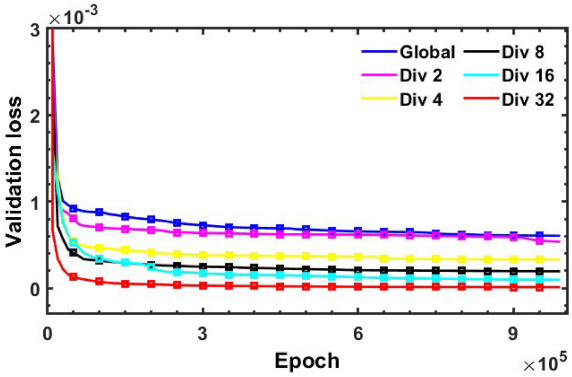

FIGURE 8. Validation loss according to the number of bins.

of the ANN model improves proportionally according to the number of bins. The optimal binning method in a practical situation is to set a small bin size in the critical regions such as the regions with the minimum geometric sizes, where the variability of the device is large and the most utilized regions. Conversely, setting a large bin size in the non-critical regions and infrequently used regions will be helpful. We recommend employing the binning method, considering careful design of experiments (DOE), physics-domain expertise, data visualization, and importance analysis.

However, if training devices cannot be extracted sufficiently, each bin must be carefully split considering the distribution and physical characteristics of the training devices. To determine whether it is wise to split each bin into new bins, we propose the decision criteria for bins using the GTBF method and hierarchical analysis. Our aim is to find the optimal bin size that maximizes the accuracy of the training devices while preserving the accuracy of the testing devices within each bin. Using the GTBF stopping condition as previously mentioned, it is possible to find the accuracy of the training devices with same accuracy of testing devices. Therefore, we use the MSE of the training devices as a metric for each bin. As shown in Fig. 9 (a), if the MSE of the ANN model trained with the separated bins is smaller than that of the ANN model trained with the original bin, the bin size is reduced. On the other hand, if the training data for the divided bins are insufficient to describe the region, the GTBF stopping condition terminates earlier than before, resulting in worse accuracy. Also, when separating bins, there may be very few test devices, in which case GTBF stopping does not work well. In the above two cases, the binning method is terminated. Finally, using Algorithm 1, the optimal bin sizes are determined as shown in Fig. 9 (b).

III. MODEL INTEGRATION AND EXPANSION

After completing the final compact model, the weights and biases of the divided ANN models are integrated into one compact model using 'if' and 'include' statements in the Verilog-A language. The integrated compact model loads the model parameters (weights and biases) according to

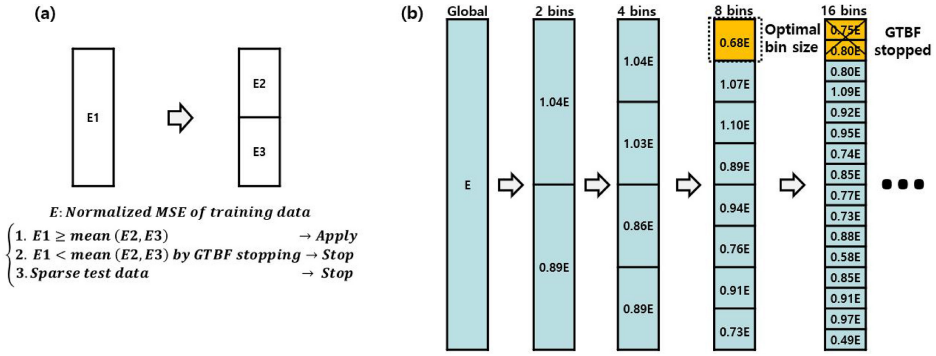

FIGURE 9. (a) Decision criteria for binning, (b) Bin size optimization flow.

Algorithm 1 Optimal Binning-ANN Model Implementation

1. **ANN Input**: geometric parameters, bias parameters
2. **ANN Output**: electrical parameters
3. Select binning parameters with high Spearman correlation coefficients
4. Define *GTBF* function
for $0 \leq i \leq T_{global}$ **do**
  Train ANN
  if i % specific interval==0 **then**
    if $MSE_{Binning}$(test) > MSE_{Global}(test) **then**
      break
    else
      continue
return $MSE_{Binning}$(train) × #(train)
5. Optimize bin sizes using *GTBF* function
    while true **do**
      if *GTBF*(a,b) < *GTBF*(a)+*GTBF*(b) **then**
        break
      else if #(test(a)) or #(test(b)) is too small **then**
        break
      else
        Divide (a,b) to (a), (b)
∗ (a,b) ≡ (a) ∪ (b)
∗ (a), (b) are data in bins
6. Complete binning-ANN

Algorithm 2 Proposed Model Integration Method Using Verilog-A

1. Define module (MOS) and terminal (D,G,S,B)
2. Define parameters (N_{fin}, I_{ds}, $C_{gg,gd,gs,gb}$ . . .)
3. Define weights, biases according to the condition
if (binning condition 1) **then**
  define weights, biases of ANN_1
else if (binning condition 2) **then**
  define weights, biases of ANN_2
. . .
else
  define weights, biases of ANN_N
4. Calculate ANN I-V model
5. Calculate ANN C-V model
6. Add electrical characteristics to the MOSFET model

binning conditions. Binning conditions can be defined in various ways, including defined geometric parameters, corner cases, and the number of stacks of NSFETs. Since the formulas used in the ANN-based compact models are the same, differently trained ANN models can be integrated into one compact model. Algorithm 2 is a verilog-A pseudo-code that forms the ANN-based compact model with multiple sets of model parameters.

The critical disadvantage of the ANN-based compact model is that it is accurate only in the range of the training devices. To use devices outside of the trained range, the model coverage of devices must be expanded by adding new data to the training data. Fig. 10 shows two different model coverage expansion methods. Fig. 10 (a) shows the conventional model coverage expansion method using only one set of weights and biases. To increase the range of geometric parameters for the ANN-based compact model, the new devices must be added to the training data and the ANN must be retrained. During retraining, a new ANN model can be designed quickly and efficiently using the transfer learning method. Transfer learning is an algorithm that can train a new model at high-speed using only a small amount of additional training data based on the weights and biases of an existing model. Transfer learning can help with the problems of large training time and insufficient training data, but it does not avoid the trade-off between the amount of training data and model accuracy. In general, in ANN models, the amount of training data has a linear relationship with the size of the model [17]. A major drawback is that the accuracy of the ANN model may be decreased by the underfitting problem as new data are added at the same model size. When the size of the model is increased to solve the underfitting problem, the degradation of simulation speed becomes a problem [7], [8]. This bottleneck reduces the reusability of the existing model and makes it difficult to expand the coverage of ANN-based compact models.

In contrast, Fig. 10 (b) shows the proposed model coverage expansion method using the binning method through the proposed Verilog-A code. In this method, the newly added

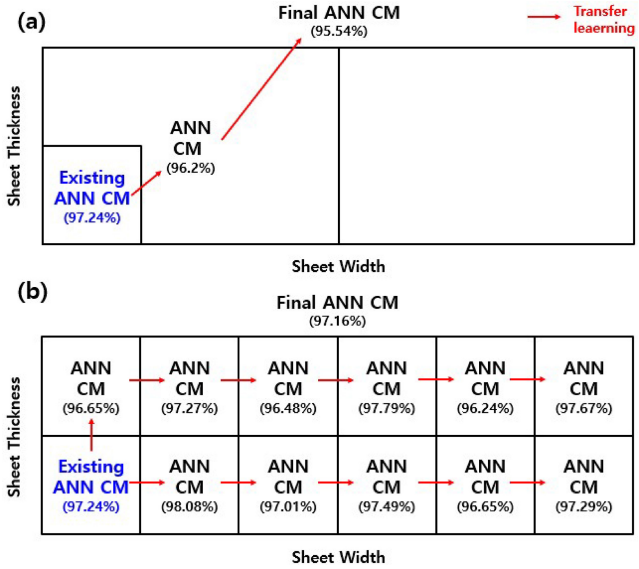

FIGURE 10. Two model coverage expansion methods. (a) The conventional model coverage expansion method using one ANN model, (b) The proposed model coverage expansion method using multiple ANN models.

TABLE 3. ASAP7 dataset.

Type	Symbol	Parameter	Range	Distribution
Geometric	L_g	Gate length	16 ∼ 26 [nm]	Sobol
Geometric	L_{sp}	Spacer length	3 ∼ 5 [nm]	Sobol
Geometric	T_{fin}	Fin thickness	5 ∼ 8 [nm]	Sobol
Geometric	H_{fin}	Fin height	16 ∼ 48 [nm]	Sobol
Geometric	EOT	Equivalent oxide thickness	0.7 ∼ 1.3 [nm]	Sobol
Bias	V_{gs}	Gate-source voltage	0 ∼ 0.65 [V]	Uniform
Bias	V_{ds}	Drain-source voltage	0 ∼ 0.65 [V]	Uniform

training data are used separately and independently from the existing model for training. The problem of reusability of the existing model is solved, and the accuracy of each model is independent. Starting from the same model, the accuracies of the two methods were compared when the range of sheet width of three-stacked NSFETs was expanded by six times and the range of sheet thickness by two times. The existing ANN model was trained with 300,000 iterations, and the other ANN models were trained with 30,000 iterations using transfer learning. In the conventional method, the accuracy of the ANN model tends to decrease as new training data are added. On the other hand, with the proposed method, even if new training data are added, the accuracy of the model is maintained because each ANN model is independently trained. This means that model expansion using the binning method is free from any restrictions like underfitting and retraining.

IV. CIRCUIT SIMULATION RESULTS AND DISCUSSION

SPICE simulation was executed to evaluate the simulation accuracy and speed of the proposed binning-ANN model. ASAP7 was used as the reference standard compact model [18]. The training dataset was extracted by Sobol sequence using HSPICE, and 1,000 devices were extracted as shown in Table 3. For eight bias conditions, $V_{gs} = 0.05$V, 0.25V, 0.325V, 0.5V, 0.65V and $V_{ds} = 0.05$V, 0.325V, 0.65V,

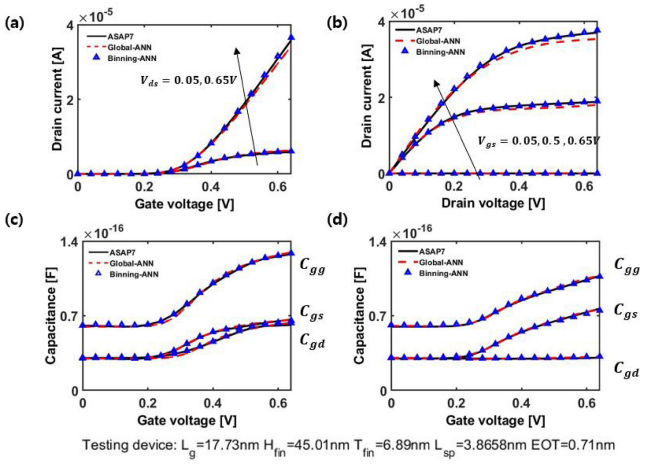

FIGURE 11. NMOS fitting results of the global-ANN and the binning-ANN. (a) $I_D - V_{GS}$, (b) $I_D - V_{DS}$, (c) $[C_{gg}, C_{gd}, C_{gs}]-V_{GS}$ at $V_{DS} = 0.05$V, and (d) $[C_{gg}, C_{gd}, C_{gs}]-V_{GS}$ at $V_{DS} = 0.65$V.

26 I-V points were extracted. For three bias conditions, $V_{ds} = 0.05$V, 0.325V, 0.65V, 26 C-V points were extracted. The total number of training data points is 1,000×26×8 = 208,000 (I-V model), 1,000×26×3 = 78,000 (C-V model). The BSIM-CMG model in ASAP7 is an analytical low-fidelity device model so that the ANN was designed with a small size because the nonlinearity and complexity of electrical characteristics according to geometric parameters are smaller than those of rigorous high-fidelity models such as TCAD [6]. To compare the accuracy and speed of the proposed model, the reference global-ANN model with the same model size and the large global-ANN model with the same accuracy were used as comparison groups. For the reference global-ANN model and the proposed binning-ANN model, the I-V model was designed with two layers, each with 5 hidden neurons, and the C-V model had one layer with 10 hidden neurons. The large global-ANN model increased the model size to have the same accuracy as the proposed binning-ANN model. The I-V model was designed with two layers, with 10 and 5 hidden neurons, and the C-V model had 1 layer with 15 hidden neurons. The binning-ANN was trained through the proposed learning method using GTBF in Section II and integrated using the model integration method in Section III. The global-ANN was converted to the Verilog-A language in the same way as in previous studies [7], [8]. Fig. 11 is the fitting result of the reference global-ANN and the proposed binning-ANN using the device that has the highest I_{on} among training devices. As shown in Table 4, the binning-ANN shows better fitting accuracy than the global-ANN at the same model size for training devices.

Two circuit testbenches were used to compare the circuit simulation speed and accuracy of each model. Circuit simulation results were verified using 100 testing devices according to the LHS uniform sampling method using HSPICE. As shown in Fig. 12, the 17-stage ring oscillator

TABLE 4. I-V, C-V fitting errors.

Error	Reference global-ANN	Large global-ANN	Proposed binning-ANN
NMOS I-V [%]	2.80	0.89	0.94
PMOS I-V [%]	2.15	0.85	0.70
NMOS C-V [%] (Cgg / Cgd / Cgs)	0.85 / 1.29 / 1.08	0.51 / 0.71 / 0.70	0.41 / 0.60 / 0.49
PMOS C-V [%] (Cgg / Cgd / Cgs)	0.94 / 1.04 / 1.41	0.43 / 0.59 / 0.62	0.47 / 0.65 / 0.65

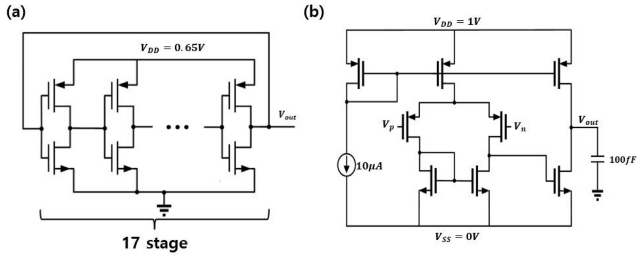

FIGURE 12. Circuit testbench schematics. (a) 17-stage ring oscillator, (b) 2-stage op amp.

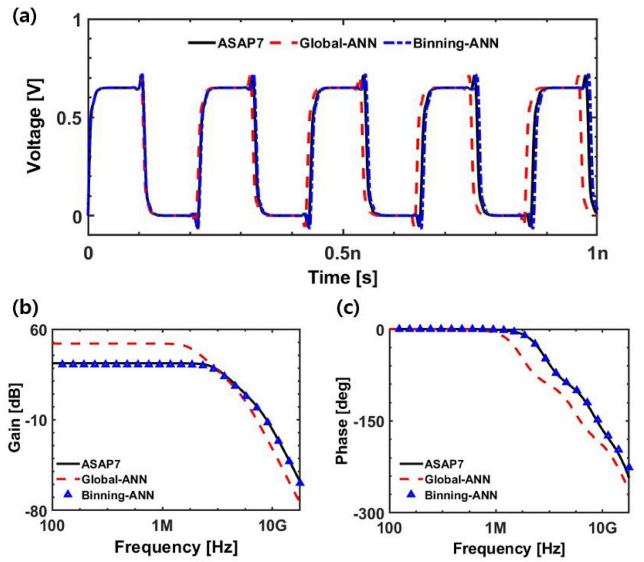

FIGURE 13. Circuit simulation waveform (a) 17-stage ring oscillator, (b) 2-stage op amp gain, and (c) 2-stage op amp phase.

(RO) is used to compare the accuracy and simulation speed of transient simulation and the 2-stage operational amplifier (op amp) is used for AC simulation. Transient simulation was performed from 0 to 1 ns using the 17-stage ring oscillator and AC simulation was performed from 100Hz to 100GHz in units of decades using the 2-stage op amp. Fig. 13 shows V_{out} waveforms of the 17-stage RO and the 2-stage op amp using the device that has the highest I_{on} among training devices. The binning-ANN shows better simulation accuracy than the global-ANN at the same model size. In addition, Table 5 shows the circuit simulation performance results according to three models for 100 testing devices. It shows that the proposed binning-ANN has higher accuracy than the reference global-ANN even on the testing devices. The proposed binning-ANN also has faster simulation speed than

TABLE 5. Circuit simulation performances.

Testbench	FoM	Reference global-ANN	Large global-ANN	Proposed binning-ANN
17-stage RO	Propagation delay Acc [%]	98.77	98.96	98.96
	Simulation time [sec] (A.u)	37.36 (1T)	50.81 (1.36T)	37.51 (1.004T)
2-stage op amp	Gain Acc [%]	74.2	93.4	95.7
	Bandwidth Acc [%]	94.45	97.18	98.63
	Phase margin Acc [%]	29.2 (Unstable)	89.2	93.46
	Simulation time [sec] (A.u)	248.39 (1T)	300.52 (1.21T)	248.81 (1.002T)

the large global-ANN. In the aspect of simulation accuracy, the binning-ANN shows up to 29% higher gain accuracy in AC simulation compared to the reference global-ANN of the same size, and shows that phase margin, which requires very high accuracy, can also be measured. In addition, it shows 36% faster simulation speed in transient simulation and 21% faster simulation speed in AC simulation compared to the large global-ANN, while maintaining better accuracy. The binning-ANN modeling method has superior simulation speed and accuracy performance compared to the global-ANN modeling method.

V. CONCLUSION AND FUTURE WORK

Existing ANN models require a larger model size to achieve high accuracy over wide model coverage. However, increasing the model size to improve the accuracy of the model reduces the simulation speed. For design space exploration and optimization, wide model coverage, high accuracy, and fast simulation speed must be guaranteed at the same time. In this study, we developed a method to achieve higher accuracy by dividing the training data into bins without increasing the size of the ANN model. Using this divided ANN model, multiple sets of model parameters are integrated into one compact model using the proposed model integration method. The proposed binning-ANN model has higher accuracy than the reference global-ANN with the same size, and has faster simulation speed and even better accuracy than the large global-ANN with the large model size. In addition, the compact model can expand model coverage with no restrictions. Because the model can be expanded independently of the existing model, a very wide range of device parameters, devices with various structures, and corner case modeling can be applied with only one compact model. Due to these advantages, the ANN-based compact model using multiple sets of model parameters will be the best option for emerging device modeling and design technology co-optimization (DTCO). As a future work, we plan to use the advantages of this model to conduct research on design space explorations and optimizations of next-generation devices.

ACKNOWLEDGMENT

The EDA tool was supported by the IC Design Education Center (IDEC), South Korea.

REFERENCES

[1] C.-T. Tung and C. Hu, "Neural network-based BSIM transistor model framework: Currents, charges, variability, and circuit simulation," *IEEE Trans. Electron Devices*, vol. 70, no. 4, pp. 2157–2160, Apr. 2023, doi: 10.1109/TED.2023.3244901.

[2] J. P. Duarte et al., "BSIM-CMG: Standard FinFET compact model for advanced circuit design," in *Proc. 41st Eur. Solid-State Circuits Conf. (ESSCIRC)*, 2015, pp. 196–201, doi: 10.1109/ESSCIRC.2015.7313862.

[3] S. Khandelwal et al., "BSIM-IMG: A compact model for ultrathin-body SOI MOSFETs with back-gate control," *IEEE Trans. Electron Devices*, vol. 59, no. 8, pp. 2019–2026, Aug. 2012, doi: 10.1109/TED.2012.2198065.

[4] S. Guglani et al., "Artificial neural network surrogate models for efficient design space exploration of 14-nm FinFETs," in *Proc. Device Res. Conf. (DRC)*, 2022, pp. 1–2, doi: 10.1109/DRC55272.2022.9855816.

[5] M.-H. Oh, M.-W. Kwon, K. Park, and B.-G. Park, "Sensitivity analysis based on neural network for optimizing device characteristics," *IEEE Electron Device Lett.*, vol. 41, no. 10, pp. 1548–1551, Oct. 2020, doi: 10.1109/LED.2020.3016119.

[6] K. Sheelvardhan, S. Guglani, M. Ehteshamuddin, S. Roy, and A. Dasgupta, "Machine learning augmented compact modeling for simultaneous improvement in computational speed and accuracy," *IEEE Trans. Electron Devices*, early access, Mar. 9, 2023, doi: 10.1109/TED.2023.3251296.

[7] S. Woo, H. Jeong, J. Choi, H. Cho, J.-T. Kong, and S. Kim, "Machine-learning-based compact modeling for sub-3-nm-node emerging transistors," *Electronics*, vol. 11, no. 17, p. 2761, 2022, doi: 10.3390/electronics11172761.

[8] H. Jeong et al., "Fast and expandable ANN-based compact model and parameter extraction for emerging transistors," *IEEE J. Electron Devices Soc.*, vol. 11, pp. 153–160, 2023, doi: 10.1109/JEDS.2023.3246477.

[9] Y. S. Cha, et al., "A novel methodology for neural compact modeling based on knowledge transfer," *Solid-State Electron.*, vol. 198, Dec. 2022, Art. no. 108450, doi: 10.1016/j.sse.2022.108450.

[10] C. Akbar, Y. Li, and W.-L. Sung, "Transfer learning approach to analyzing the work function fluctuation of gate-all-around silicon nanofin field-effect transistors," *Comput. Electr. Eng.*, vol. 103, Oct. 2022, Art. no. 108392, doi: 10.1016/j.compeleceng.2022.108392.

[11] Y. Kim et al., "The efficient DTCO compact modeling solutions to improve MHC and reduce TAT," in *Proc. Int. Conf. Simul. Semicond. Process. Devices (SISPAD)*, 2018, pp. 58–61, doi: 10.1109/SISPAD.2018.8551725.

[12] C. C. Tan and P. B. Y. Tan, "Accurate BSIM4 MOS model extraction with binning-hybrid-macro methodology," in *Proc. IEEE Int. Conf. Semicond. Electron. (ICSE)*, 2016, pp. 89–92, doi: 10.1109/SMELEC.2016.7573598.

[13] H.-L. Chang, Y. Ma, and Z. Liu, "Enabling efficient design-technology interaction by spec-driven extraction flow," in *Proc. IEEE Int. Electron Devices Meet. (IEDM)*, 2020, pp. 41.5.1–41.5.4, doi: 10.1109/IEDM13553.2020.9372118.

[14] H. Kang, Y. Wu, L. Chen, and X. Zhang, "Research on device modeling technique based on MLP neural network for model parameter extraction," *Appl. Sci.*, vol. 12, no. 3, p. 1357, Jan. 2022, doi: 10.3390/app12031357.

[15] J. Wei, H. Wang, T. Zhao, Y.-L. Jiang, and J. Wan, "A new compact MOSFET model based on artificial neural network with unique data preprocessing and sampling techniques," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 42, no. 4, pp. 1250–1254, Apr. 2023, doi: 10.1109/TCAD.2022.3193330.

[16] I. M. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR Comput. Math. Math. Phys.*, vol. 7, no. 4, pp. 86–112, 1967, doi: 10.1016/0041-5553(67)90144-9.

[17] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Int. Conf. Artif Neural Netw.*, 2018, pp. 270–279, doi: 10.1007/978-3-030-01424-7_27.

[18] L. Clark et al., "ASAP7: A 7-nm finFET predictive process design kit," *Microelectron. J.*, vol. 53, pp. 105–115, Jul. 2016, doi: 10.1016/j.mejo.2016.04.006.