Ph. D. DISSERTATION

Machine Learning Strategy for Predicting Process Variability Effect in Ultra-scaled GAA FET and 3D NAND Flash Devices

초소형 GAA FET 소자 및 3D NAND Flash 소자의 공정 변동성 영향을 예측하기 위한 기계 학습 전략

BY

Kyul Ko

August 2020

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Machine Learning Strategy for Predicting Process Variability Effect in Ultra-scaled GAA FET and 3D NAND Flash Devices

초소형 GAA FET 소자 및 3D NAND Flash 소자의 공정 변동성 영향을 예측하기 위한 기계 학습 전략

지도 교수 신 형 철

이 논문을 공학박사 학위논문으로 제출함

2020 년 8 월

서울대학교 대학원

전기정보공학부

고 결

고결의 공학박사 학위논문을 인준함

2020 년 8 월

위 원 장 최 우 석 (인)

부위원장 신 형 철 (인)

위 원 김 장 우 (인)

위 원 강 명 곤 (인)

위 원 김 윤 (인)

ABSTRACT

This paper presents a Machine Learning (ML) approach for accurately predicting the effects of various process variation sources on ultra-scaled GAA FET devices and 3D NAND Flash Memories. The effects of process variability sources cause various reliability problems in logic and memory devices. In particular, accurate prediction and control is essential by reducing the margin that determine the yield of logic and memory devices.

The machine learning system is largely divided into three classes: Unsupervised Learning, Supervised Learning, and Reinforcement Learning. Among them, a supervised learning series machine learning system, which uses a regression method to train predictive models based on input and output (Training data) values, is the most suitable method for analyzing device characteristics and predicting variability effects. Since the machine learning system of the supervised learning series needs to predict the characteristics of various devices of various variability sources, it is possible to use multiple input-multiple outputs (MIMO) based on complex algorithms (artificial neural networks) with multiple nodes (MN).

In the early stages of the ML system, the variability sources of a single transistor is analyzed. We propose an accurate and efficient machine learning approach which predicts variations in key electrical parameters using process variations (PV) from ultra-scaled gate-all-around (GAA) vertical FET (VFET) devices. The proposed machine learning approach shows the same accuracy and good efficiency when compared to 3D stochastic Technology-CAD (TCAD) simulation. Artificial Neural Network Based (ANN) machine

learning algorithm can perform Multi-input-Multi-Output prediction very effectively.

As an advanced stage of the ML system, we propose a variability-aware ML approach that predicts variations in the key electrical parameters of 3D NAND Flash memories. For the first time, we have verified the accuracy, efficiency, and generality of the predictive impact factor effects of ANN algorithm-based ML systems. ANN-based ML algorithms can be very effective in MIMO prediction. Therefore, changes in the key electrical characteristics of the device caused by various sources of variability are simultaneously and integrally predicted. This algorithm benchmarks 3D stochastic TCAD simulation, showing a prediction error rate of less than 1% as well as a calculation cost reduction of over 80%. In addition, the generality of the algorithm is confirmed by predicting the operating characteristics of the 3D NAND Flash memory with various structural conditions as the number of layers increases.

Keywords : Process Variation (PV), Machine learning (ML), Artificial neural network (ANN), Gate-all-around (GAA), Vertical device, NAND Flash Memories, Prediction.

Student number : 2015-20882

CONTENTS

Chapter 1
Introduction

1.1 Emergence of Ultra-scaled 3D Device

Technology shrink to improve the performance and integration of MOSFET devices has been the surest way of developing next-generation technologies. During the last four decades, the semiconductor industry has been striving to develop technology for device scaling and meeting Moore's Law. Although there are many opinions that the limit of reduction is approaching, the current semiconductor technology has reached the semiconductor device of 10 nm or less. According to the latest ITRS roadmap, half the pitch of logic FinFET is expected to shrink to 15nm and 13.4nm in 2020, and 3D NAND flash memory is expected to develop the same level of technology.

Figure 1.1 shows the development through the introduction of high-K/metal gate process and strain engineering technology for logic devices. Starting at the 20nm technology node, planar MOSFETs were first transformed into 3D FinFET transistor technology, which continues to develop Moore's Law in terms of area and power and speed.

More than 20 foundaries have been produced with 130nm technology due to the explosive increase in fab and chip design costs, but 5 major companies for processes of 10-7nm and above may be announced. According to the announcement of the world's five leading groups, test chips below the 7nm technology node have been announced so far,

and the device structure after FinFET will be provided under the 5nm technology node in the future.

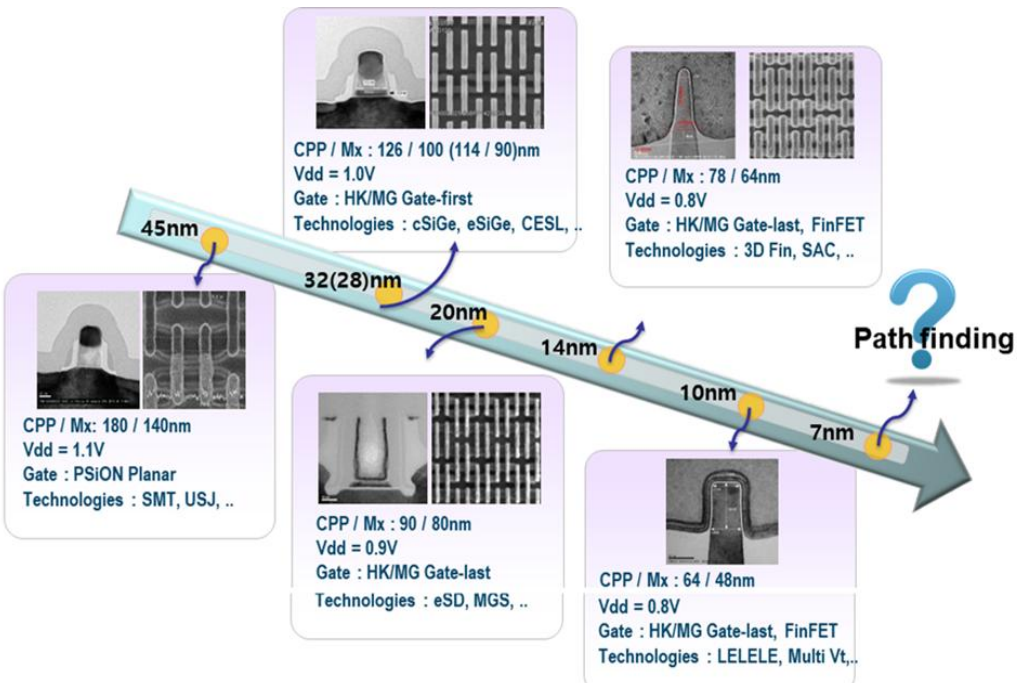


Fig. 1.1. Logic transistor development direction.

In the case of logic devices, new 3D structures or materials are under active research in next generation semiconductor nodes below 10 nm, but have not yet been decided. Especially for logic devices, gate-all-round multi-nanowires and multi-nanoplates (sheets) are considered at 7nm / 5nm. At the Samsung Foundry Forum held in Santa Clara, USA on May 24, 2017, Samsung Electronics announced plans to apply multi-bridge-channel

transistors (MBCFETs) to 4nm technology nodes.

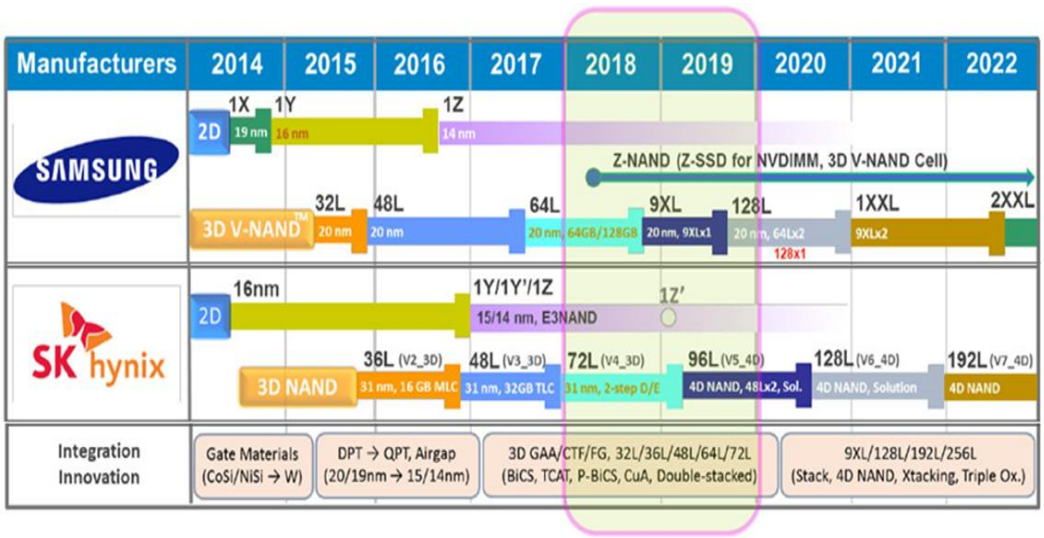


Fig. 1.2. 3D NAND Flash memories development direction.

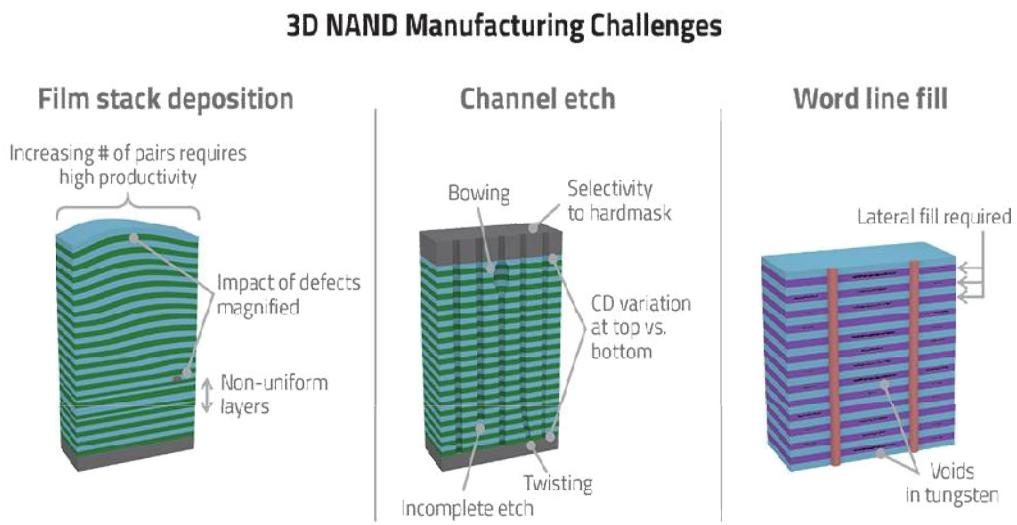


Fig. 1.3. 3D NAND Flash memories Variability issues.

Figure 1.2 shows that while 3D NAND flash memory has evolved to a higher stacking layer for integration, the device size has decreased significantly. According to the announcement of the world leader in 3D NAND flash memory, more than 192 vertical stacks will be targeted in the future, resulting in smaller device sizes. Therefore, devices with various structures are proposed, but the effects of variable sources are expected to be intensified.

Due to high integration of memory cells, NAND flash memory has evolved into a 3D NAND structure with a planar structure. 3D NAND structures can be modified by a variety of factors, including grain boundary trap (GBT) effects in polysilicon, critical dimension control (CD Control) effects, and taper angle effects. It is shown in Figure 1.3.

1.2 Increasing Difficulty of Interpreting Variability Issues

In real life, many electronic products are used to assess the ability of a process to perform its intended task, taking into account the sources of variability for each application area. Electronic product suppliers must ensure that their products operate within acceptable limits.

If we can't guarantee the normal operation within the acceptable range, after product development, not only will you incur additional development costs like new development, but also negatively affect the customer's product image and brand image.

Figure 1.4 shows an example of a logical device and shows the effect of various variability sources as the technology node decreases. Relatively large device considers only major sources of variation due to the limitations of each process technology. However, as the size of devices continues to decrease, further major sources of variability need not be considered, but all variability factors need to be comprehensively predicted.

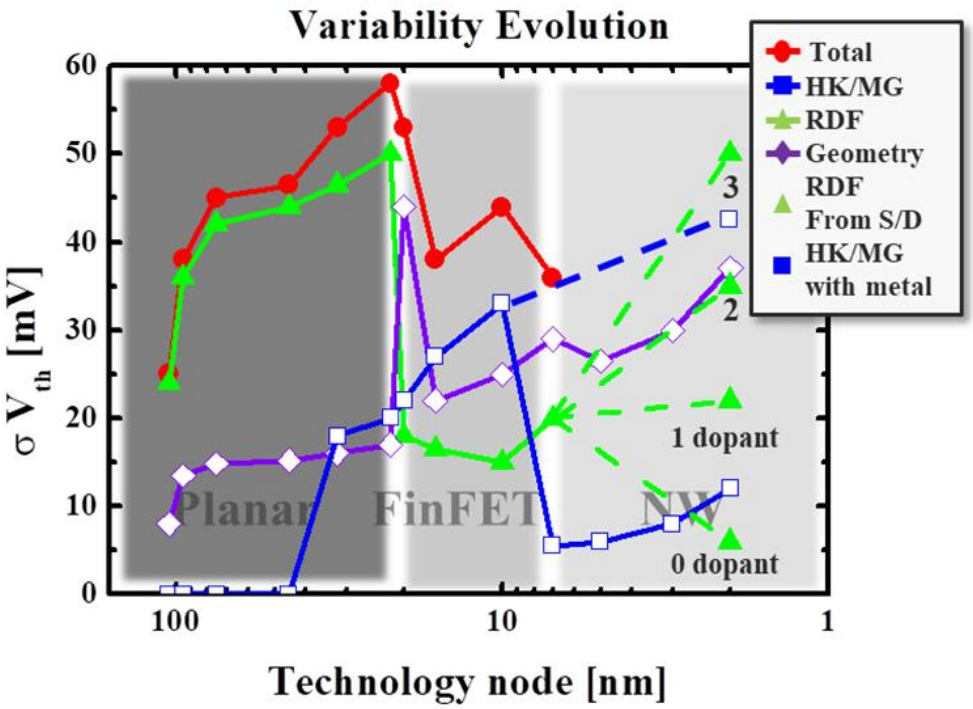


Fig. 1.4. Variability sources impact of technology node reduction.

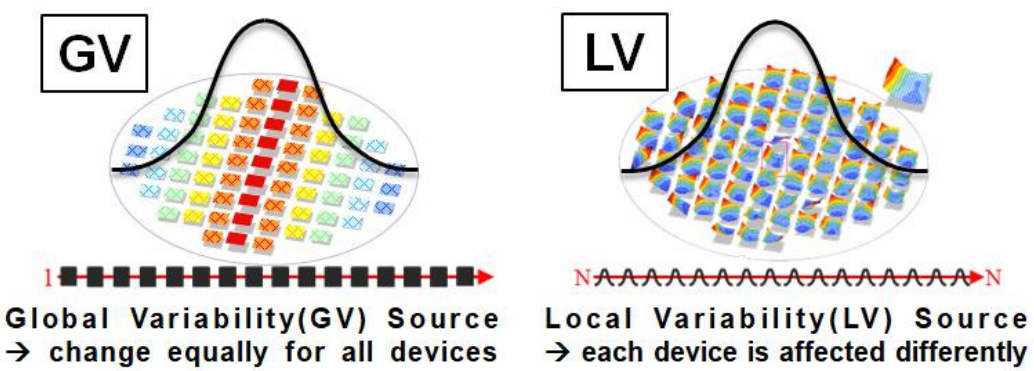


Fig. 1.5. Global variability sources and local variability souces schematic.

Figure 1.5 shows the mechanism that causes variable problems in next-generation semiconductor device technology. It is divided into local variability sources (LV) and global variability sources (GV). The LV source is the variability sources that occur within a single device, and typically includes line edge roughness (LER), work function variation (WFV), random dopant fluctuation (RDF), and grain boundary trap charges (GBT). The GV source is caused by chaneges in physical variables such as gate length, channel thickness and equivalent oxide thickness based on CD (Critical Dimension) control.

Local and global variations are due to the process technology limitations and the intrinsic properties of materials. Therefore, there is a problem that the characteristics of the device are fatally affected by the rapid decrease of the technology nodes. The effects of these various sources of variation change various electrical characteristics of the device, and common related attributes include transistor threshold voltage (Vth), operating current (= current, ion or saturation current, Isat), Leakage current. (= Off current, Ioff), subthreshold swing (= subthreshold swing, SS) and transconductance (gm) are considered. As a result, the performance of a single device is degraded and a 3D NAND flash memory failure occurs.

The application of low power supply voltage due to physical area reduction and scale down increases the variation, and the use of various 3D structures and new materials aiming for higher integration makes the variation analysis more difficult.

In 3D structures, a variety of GV and LV sources have emerged that are largely

unconsidered from traditional planar structures, which also affects the variability mechanism of the device. Also, these various variability sources interact with each other, no longer ignoring the effects of variability sources at the atomic level. The emergence of these various variability factors increases the overall complexity of variability prediction.

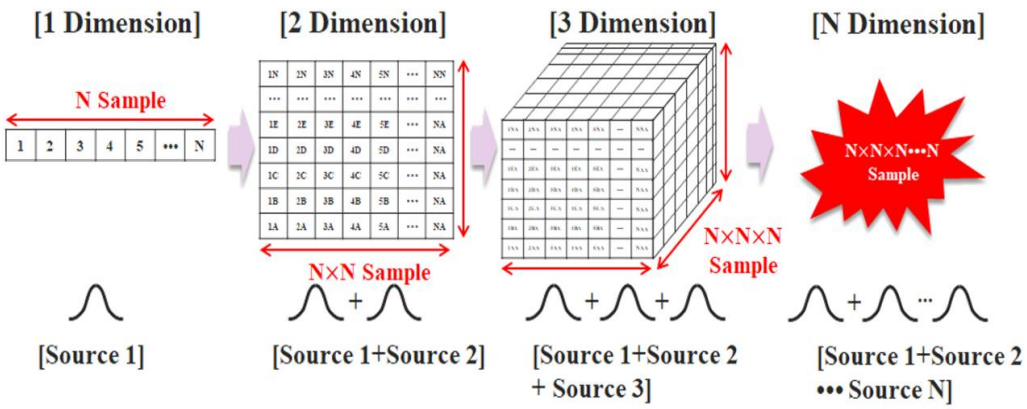


Fig. 1.6. Increasing computational costs incurred by simultaneously considering various variability sources.

Figure 1.6 shows the increase in the minimum number of predicted samples as the number of variable sources analyzed increases. In situations where different variability sources are expected at the same time, more sample devices are needed than before, increasing computational cost and analysis difficulty.

In short, scaling down and 3D structure deepens the effect of variability and increases the complexity of variability (three-dimensional structure, atomization effect, mechanical

effect, variation effect) analysis, so accurate variability analysis is very important for the development of next-generation processes.

The emergence of 3D devices considers not only the effects due to the structural properties of the device, but also the physical effects that may occur in the ultra-scaled devices. Therefore, a new system is needed that accurately predicts the causes of the various variations that result from these effects. Traditional TCAD simulation approaches have limited the variability of source analysis methods due to their limited and inefficient computational costs. However, in situations where analytical methods through TCAD simulation are commonly used, the method described in this study is a new system through machine learning. This advanced variation analysis system overcomes the limitations of existing TCAD simulations and allows new systems to be used without the additional overhead.

1.3 Need for Accurate Variability Prediction

The device architecture, dimensions, materials, etc. are set through PPA (power = power consumption, performance, = speed, area = integrated screen) analysis and prediction in the early stages of next-generation semiconductor device development. Variability is usually studied at the time of hardware release. However, as variability issues increase, it is necessary to predict variability early in development and determine operational and overdrive bias in order to reduce development time and reduce development risk hedging. Therefore, prediction through TCAD simulation has been actively used for the technology definition of the next-generation semiconductor development node in the early stage of development node. It is mainly evaluated based on Figure of Merits (= FOMs: Ion, Ioff, Vth, SS, gm).

Due to the recent diversification of variability sources and the increasing complexity of prediction, accurate prediction through TCAD simulation has become even more important. Therefore, analysis of large sample devices through the traditional variability method, the Statistical Impedance Field Method (sIFM), allows accurate integrated assessment of variability sources, but is computationally expensive. As a result, there are limits in terms of analytical efficiency.

A commercial tool for analyzing the effects of variability sources (= sIFM-based TCAD Sentaurus simulator) sets Poisson's equation and current transfer by setting direct and indirect variables that determine each variability sources as inputs. Returns the

electrical characteristics of a large number of sample devices calculated by the function. From the perspective of analyzing the effects of variability, the more sample device output data, the more accurate the analysis. However, analysis of a large number of sample devices is computationally expensive, so it is important to set the optimum number of samples. Unfortunately, the number of samples analyzed should be greater than or equal to the minimum requirement, as analyzes of different variability sources need to consider the effects of each other.

The changes in electrical characteristics due to individual variability sources are based on previous research results. For example, Figure 1.7 shows that for a Nanoplate (Nanosheet) VFET under a 5nm technology node that includes local and global variability sources of variation, the effective standard deviation of the threshold voltage variation is about 19mV. Figure 1.8 shows the standard deviation level of the threshold voltage variation distribution for 3D NAND flash memory with both local and global variability sources. Therefore, when determining the minimum number of valid data based on the number of target variability sources, it is possible to predict and determine the influence of variability with an error rate of 1% or less when the minimum sample data exists according to the number of variability sources.

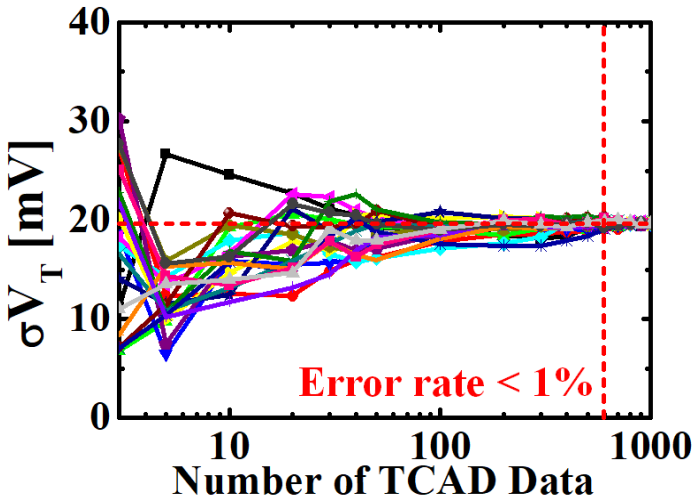


Fig. 1.7. Example of effective variation analysis based on the number of sample data of logic devices.

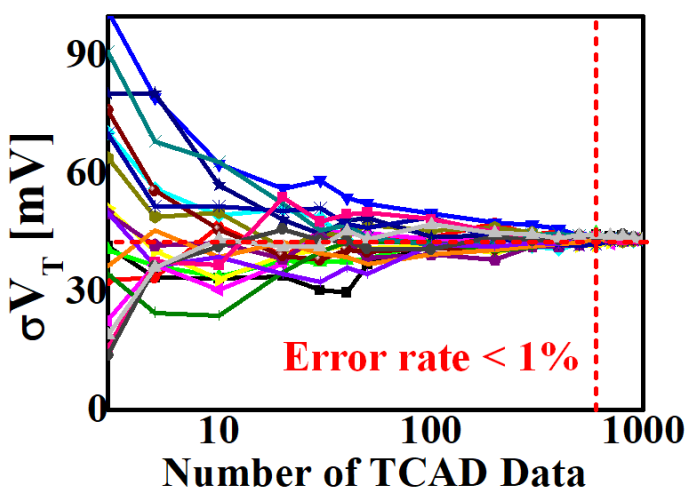


Fig. 1.8. Example of effective variation analysis based on the number of sample data of 3D NAND Flash memories.

When targeting the ultra-scaled high stacked device, it is necessary to analyze the integrated effects of various variability sources. Therefore, to analyze an integrated variable source instead of a single variable source requires research based on commercial tools to set the minimum number of valid data. The setting for the minimum number of valid data can be increased according to the number of sources of fluctuation, with the goal of calculating the lowest calculation cost with an error rate of less than 1% in priority.

Variation source prediction in next-generation semiconductors through TCAD is a highly integrated method with accuracy, but at the same time it consumes high computational costs. Therefore, according to a number of previous studies, a method of predicting the cause of each variation using a compact model has recently been proposed as a method for increasing the analysis efficiency. Compared to the existing TCAD method, the compact model is significantly faster than the existing TCAD method, but there are officially large errors due to assumptions and approximations, and the integrated analysis method is limited. Therefore, by taking advantage of TCAD's superior accuracy and integrated analysis, it is possible to reduce the computational cost by training various extracted data with the ML system of the supervised learning series. As a result, the ML system that uses big data (Technology-CAD-based data) as the source of variation analysis can predict the analysis of various variability sources efficiently and accuratel.

Through process optimization, variability can be sufficiently overcome in next-

generation semiconductor development nodes and, in some cases, better optimization in terms of variability than previous generations. Therefore, optimization by integrated variability simulation is essential for process development.

In conclusion, this study develops major variability mechanism physics-based model and simulator-based analysis platform that including local variability sources and global variability sources and applies it to big data-based ML systems. A complete system of implementation can be presented to contribute to process optimization. Therefore, this study aims to suggest a platform that can be utilized in the industry that can find the optimal technology at the early stage of development in terms of variability in the next generation semiconductor technology.

Chapter 2
Machine Learning System

2.1 Introduction

The machine learning system is divided into three major classes: Unsupervised Learning, Supervised Learning, and Reinforcement Learning [1]. Among them is a supervised learning series ML system that uses regression methods to analyze device characteristics and train predictive models based on input / output (training data) data to predict the effects of variations. This is the most suitable method for this research. The ML system of the supervised learning series aims to predict the device characteristics of devices at different angles of various variability sources. It can also be developed via MIMO based on complex algorithms with MN (eg artificial neural networks, ANN). The implementation framework of ML system can be configured based on MATLAB and Python, and given the comprehensive problems, the best framework for this research is used for the system through Python series Pytorch (See Table 2.1). Also, as shown in Table 2.2 comparing various ML algorithms, the most suitable method for analyzing the variable source of a semiconductor device is determined by ANN's algorithm.

Tool	Framework	Key features
MATLAB	Statistics and Machine Learning Toolbox (Mathwork Toolbox)	➢ Higher mathematical precision than other tools using MATLAB ➢ Requires low code compatibility and deep MATLAB knowledge ➢ Library is not diverse because there are not many users at present
python	TensorFlow by Google	➢ Open source, user friendly, relatively short time to learn ➢ Advantageous for projects that perform complex workflow slowly (relative time for model formation and configuration) ➢ Support the largest amount of libraries
	PYTORCH Deep Learning with PyTorch by facebook	➢ Open source, user friendly, relatively short time to learn ➢ Favorable project to check and correct results within a short time (short time due to simple and intuitive model formation) ➢ Large library support ➢ Suitable for this study in consideration of comprehensive matters

Table. 2.1. Compare machine learning commercial tools.

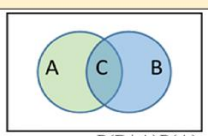
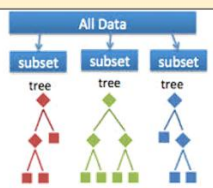
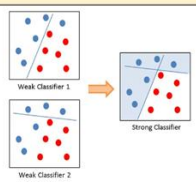
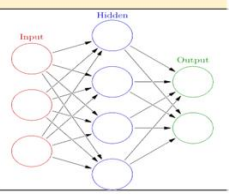
	Naive Bayes	Random Forests	AdaBoost	Neural Networks
Schematic	 $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$			
Average Prediction Accuracy	Low	High	High	Very High
Learning speed	High	Low	Low	Medium
Predicting speed	High	Medium	High	High
Nonlinear learning ability	Possible	Possible	Possible	Possible
Characteristics	- Complete Probability Based Algorithm - No regression analysis	- Model of Decision Trees - Possible error in successive inputs	- Select only essential properties, boost by reducing dimensions - Vulnerable to Noise and Outliner Learning	- Highest accuracy model for implementing deep learning - Most used for current complex model analysis

Table. 2.2. Compare machine learning Algorithms.

2.2 Analysis of Variability through TCAD Simulation

The goal of this study is to develop a predictive physics-based model and analysis platform that can delineate the dynamic characteristics of the major variability mechanisms (local and global variability) in next-generation semiconductor devices. It implements an ML system and develops a variability platform available in the industry. We also propose a predictive model that utilizes the developed platform to change the various sizes of next-generation GAA FET devices and ultra-high-rise 3D NAND flash memory devices to next-generation semiconductors to find optimization processes. The optimum process criterion is a quality level that takes into account the electrical characteristics and variability of the benchmark single logic device (NW VFET, NP VFET) and memory device (3D NAND flash). Therefore, the TCAD simulation environment for linking with the ML system establishes a statistical variability prediction environment by developing the following three important key technologies [2, 3].

In the [Identify] phase, we build a commercial tool-based 3D TCAD simulation and a large model-based simulation environment to analyze independent, detailed variability sources.

In the [Analysis] phase, to analyze device level effects based on extracted and predicted sources of variability, we have developed an analytical model and simulator

with traditional approach supported by TCAD simulation and large-scale data sampling.

In the [Interlock] phase, Simultaneous development of device operating characteristics and variability prediction systems by interoperating TCAD simulation and ML from a single device to a 3D NAND flash memory device.

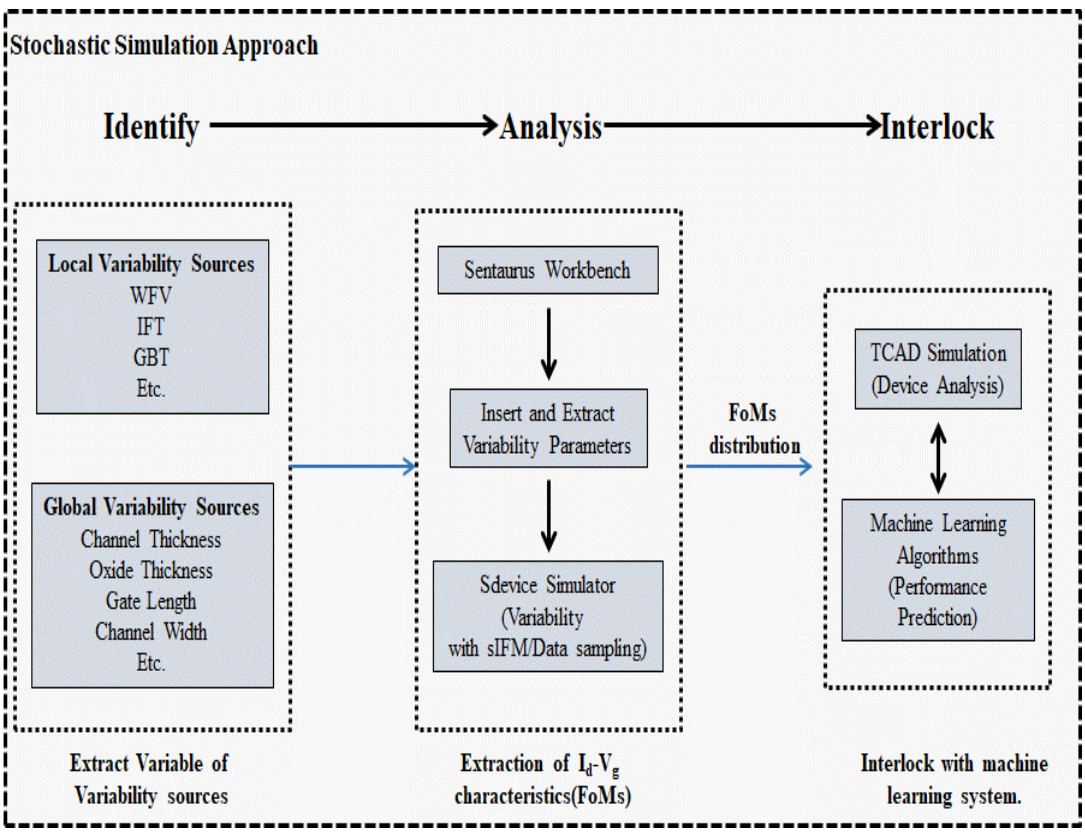


Fig. 2.1. The next-generation semiconductor variability analysis platform. Schematic diagram of detailed analysis and prediction strategies for each step.

2.2.1 Identify Phase

In the Identify phase, we build a commercial tool-based 3D TCAD simulation and a large model-based simulation environment to create independent, detailed, variable-source analysis at the atomic level. To implement an ML system that works with commercial TCAD tools covered in this paper, we first need to understand how to sample large-scale data based on the sIFM-based TCAD Sentaurus simulator and Atomistic approach.

In general, the influence of variability were analyzed in various previous studies by 3D TCAD simulation. The variability analysis of this TCAD simulation is based on the Statistical Impedance Field method (sIFM). The impedance field matching method (IFM) was proposed by Mayergoyz and Andrei(DOI: 10.1063/1.1390499) and then modified by A. Wettstein et al, SISPAD, 2003(DOI: 10.1109/SISPAD.2003.1233645). Instead of solving the entire nonlinear Poisson and drift-diffusion equations for large samples of random devices, the IFM method treats randomness as a perturbation of one reference device. Then calculate the current variations at the terminals of the device due to any of these disturbances. The IFM uses noise modeling for analysis of variability and can be used in the Sentaurus 3D-TCAD Device simulator. The IFM splits the fluctuation analysis into two tasks. The first task is to provide models for local microscopic fluctuations inside the devices. At this stage, various distribution types for variability sources are physically implemented. The second task is to determine the impact of the local fluctuations on the

terminal characteristics. To solve this task, the response of the contact voltage or contact current to local fluctuation is linear. For each contact, Green's functions are computed that describe this linear relationship. Unlike the first task, the second operation is a pure number as it is performed by the physical model specified for the numerical simulation. In addition, it is possible to calculate according to the change of the physical model through it, but it is a method which consumes expensive calculation cost [3].

The TCAD tool-based analysis of variability sources is a large number of samples calculated using Poisson equations and current transfer function, setting the inputs for direct and indirect variables that determine individual variability sources as inputs, and then returns the electrical characteristics. However, the process of analyzing mutable sources with TCAD simulations has the problems of complex 3D structure convergence and long analysis times due to the large number of sample device-based physical models that need to be calculated. To solve this, even using a simple structure, additional consistency issues may arise from the inconsistency of the structure. Nevertheless, TCAD simulations can still accurately predict the mechanisms of various variability sources that require atomic level analysis for this study, and to extract analytical variables, which are linked with the ML system for integrated variability analysis in device analysis. In addition, the TCAD tool's variability analysis system is set up with separate independent / correlated variability sources to quantitatively assess the minimum computational cost.

2.2.2 Analysis Phase

In the [Analysis] step, to analyze device level effects through extracted and predicted sources of variability, we use an analysis approach (sIFM) supported by TCAD simulation and analytical models and simulators with large data sampling [4].

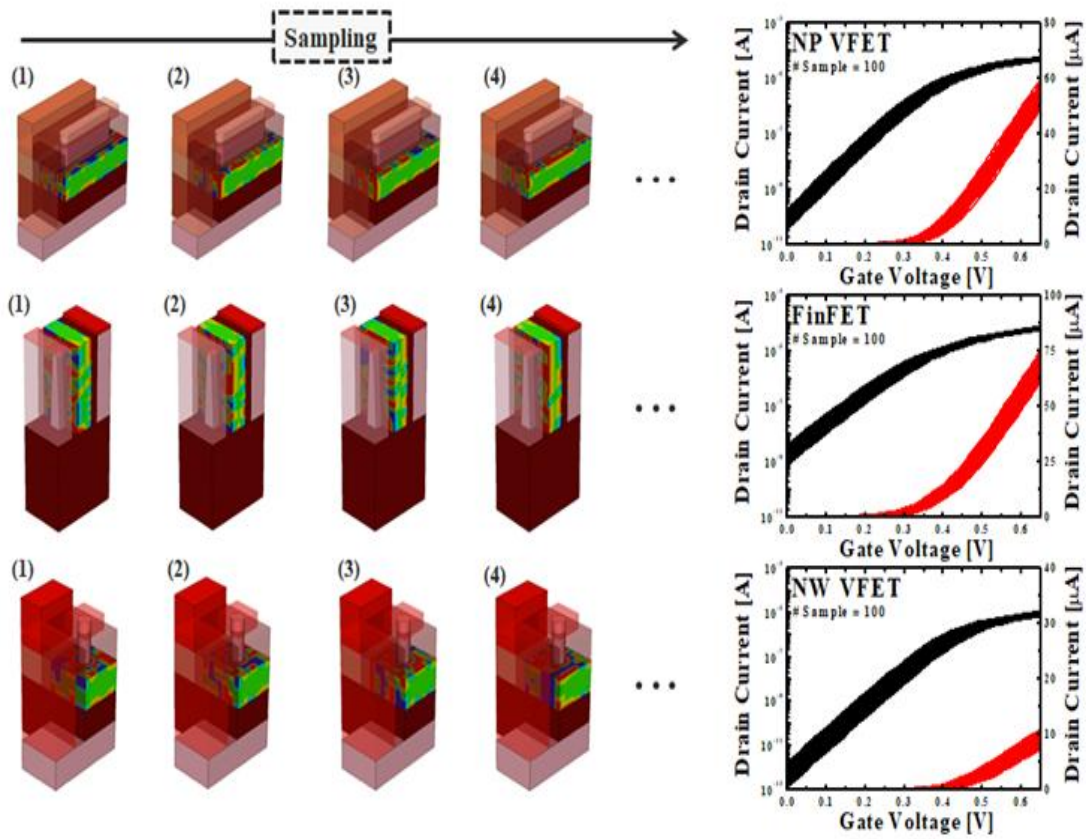


Fig. 2.2. Example of analysis of variability souces in TCAD Simulation with large scale sampling.

It is necessary to analyze the change in the electrical characteristics of the device that occur based on the extracted and predicted input variables of various variability sources. As the device size decreases, it is necessary to analyze accurately the variable source that fully considers the characteristics of single logic devices and high stack memory devices with a three-dimensional structure, and related factors are additionally considered. Therefore, unlike the existing method of analyzing variability sources, the need for more advanced and higher-level analysis methods is noted.

The variability source can be classified into two categories as follows. First, Local Variability sources (LV). Second, Global Variability sources (GV). LV sources are variability sources that occur within a single device, typically line edge roughness (LER), work function variation (WFV), random dopant fluctuation (RDF), interface trap charges (ITCs), and grain boundary trap charges (GBTs). GV sources are variability sources that occur in the entire wafer and system, and are caused by the variation of physical variables such as gate length, channel thickness, and equivalent oxide thickness based on critical dimension (CD) control.

Various variability sources included in the LV source make it difficult to extract the mutable parameters explicitly. The variable approach contained in the LV source is displayed as probability data, so the mathematical approach does not guarantee accurate predictions. For example, each local variation parameter includes variation amplitude (= Amplitude, Δ), correlation (= Correlated length, λ), grain size (Gs), number of total grain (Ng), average trap concentration (Nt), etc. It will occur locally within a single logic

device. Therefore, it is necessary to specify the direct and indirect parameters of these local occurrence to preprocess the data that can be analyzed. This study also set up an environment in which existing local change parameter data could be preprocessed as treatable data rather than unpredictable values.

GV sources are known to fluctuate device constants (e.g., gate length, channel thickness, oxide thickness, etc.), due to process limitations and specific material properties. The GV source refers to the global influence of the wafer-wafer phase and, unlike the local variability source, can be directly predicted into a variable based on CD control estimate. Therefore, in this study, we aim to quantify the direct parameters of global variability by clearly identifying integrative correlation between data preprocessing and integration of LV parameters. Therefore, this process is used to build a preprocessing environment as data for inserting integrated analytical data from various sources into the ML system.

This study extracts the training data that can be applied directly to the ML systems using device-level feature results extracted with TCAD-based analysis approach. Therefore, by using the preprocessed data to construct a variation analysis system, it is possible to develop an analysis simulator with high accuracy, integration, and calculation efficiency.

2.2.3 Interlock Phase

In the interlock phase, TCAD simulation is used to simultaneously optimize device-level operating characteristics and variability. Therefore, TCAD-based analysis tools are considered to be the most powerful and highly accurate analysis tools for assessing device level variations, but there is no way to overcome the high computational cost problem. Recently, there has been active research on improving analytical efficiency through compact models, but the aspect of integration with lower accuracy compared to TCAD-based analytical tools has become a problem.

In other words, using a commercially available 3D TCAD tool with the variability analysis method at the stage described above, after characterization of the device, the ML system is trained to link the platform to predict the variability source. Moreover, the electrical characteristics of the device can be confirmed through the SPICE simulator (HSpice, SmartSpice) based on BSIM-CMG model according to DTCO flow as well as TCAD simulation. The integration of these various tools can improve the device development efficiency and save many indirect costs. In addition, the variability mechanism simulator through ML can periodically checks the characteristics linked with commercial available TCAD tools to analyze and optimize the characteristics and variability of the operating device at the same time.

2.3 Structure of Machine Learning Algorithm

The optimal system environment is based on various input / output training data extracted from TCAD simulation, which is the core of the entire variability analysis system, to build an optimization method for the ML characteristics of supervised learning, set to the following steps [1, 5, 6].

1. In the early stages, we need to build a system that can process large amounts of training data. Therefore, the system is configured using a MN algorithm of MIMO.

2. Train the data extracted from the TCAD simulation on the ML system algorithm constructed in the previous step.

3. The training data can be trained iteratively using the internal model of the built-in ML, and in the process, the error of the ML prediction value and the training data value is checked whether the algorithm is optimized inside the ML.

4. After optimizing the algorithm, determine the sufficiency of the extracted training data and proceed to the next analysis step.

5. When it is judged that the algorithm of the ML system has been trained correctly through sufficient training data, the correlation and influence with each variability source are rapidly predicted as characteristic changes due to various parameters.

The method of analyzing variation through the ML approach introduced in this paper uses the ANN algorithm of the supervised learning series. The basic form of the algorithm is shown in the figure below:

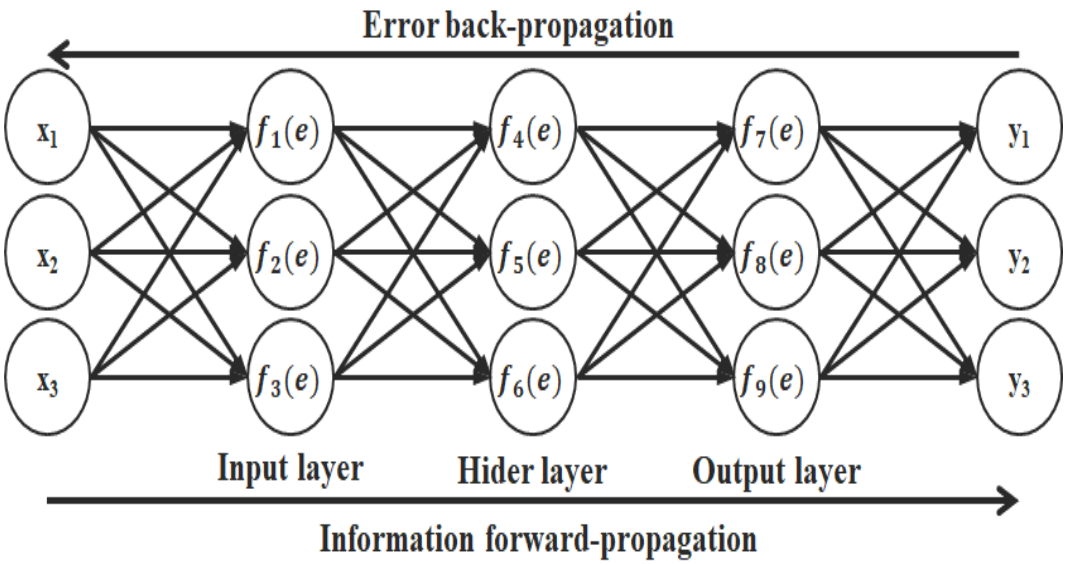


Fig. 2.3. Schematic diagram of a basic artificial neural network algorithm structure.

To effectively train the effects of variability sources on the structure of these basic algorithms, four factors need to be optimized:

1) Number of input data/output data: Set the number of data according to the number of variability parameters(Input data) and device parameter(output data). In this paper, both input data(LV and GV sources) and output data(electrical characteristics) are set.

2) Depth/Number of hidden layer: For optimization of the algorithm, set the number of

layers (e.g., input layer, hidden layer, and output layer) and the number of node of each layer (e.g.,f1(e), f2(e), f3(e) etc.). In addition, this part can control the degree of training depending on the number of layers and the number of nodes, and excessive layer number and node can cause overfitting problem of ANN algorithm. As the number of layers and nodes increases, the complexity of the algorithm model increases, overfitting and directly contributing to the problem, as shown in the figure below [1, 7].

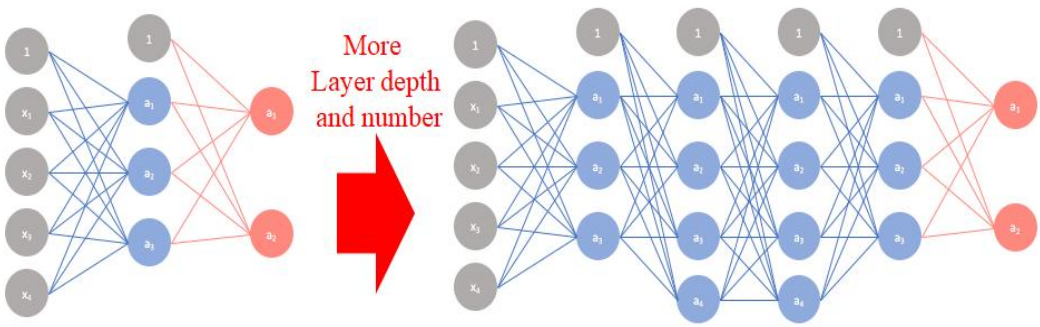


Fig. 2.4. Example of algorithm inside compute node.

On the contrary, if the number of layers and nodes are set smaller than the required values, there will be many errors in the ML system as the underfitting problem will occur and ANN algorithm will not be able to train sufficiently.

As shown in the following graph, the underfitting problem and the overfitting problem may occur according to the number of layers and nodes of the ANN. Therefore, an optimized model structure is required and a number of optimized iterative training is required.

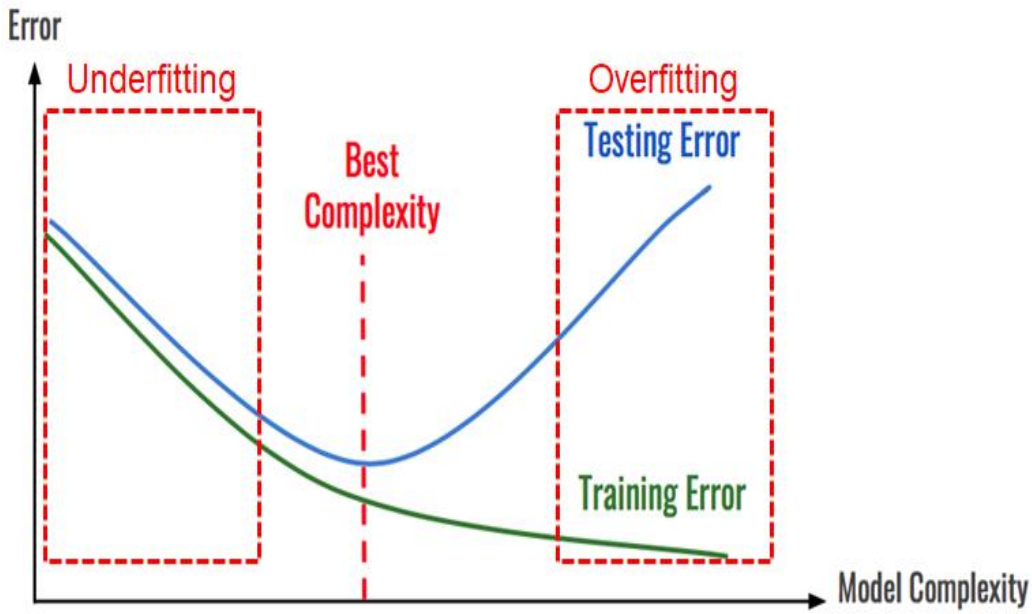


Fig. 2.5. Error rate change according to algorithm complexity. Optimize model complexity by comparing underfitting and overfitting conditions.

Therefore, the number of layers and nodes in 1: 3: 1 (Input layer: Hidden layer: Output layer) proposed in this paper are set considering the optimal complexity without problems of underfitting and overfitting.

3) Forward-propagation model: In this paper, forward learning is performed using the Rectified Linear Unit (ReLU (x) = max (0, x)) model.

4) Back-propagation model: Calculate the mean squared error (MSE) loss between the model's output data and the TCAD simulation data, and use the optimized model with the

stochastic gradient descent (SGD) method and The Adaptive Moment Estimation (ADAM) method.

Based on the above conditions, our group can design a basic ANN algorithm structure and use the data extracted through TCAD simulation to train the ML system. Figure 2.6 shows an example analysis of the effects of variability effect used in this study. The setup algorithm shows the process by which the variability source variable data entered through the activation function is simultaneously propagated to each next node and back to the previous node through the inverse correction function through error calculation. Inverse correction process during one epoch in this paper, the number of iterations (epoch) is repeated more than 20,000 times. Therefore, the error between the data output through the ANN algorithm(Y1: YN) and the determined TCAD simulation data(G1: GN) is calculated through the MSELOSS function in each forward and backward training process repeated during the training process. As a result, iterative training of the ANN algorithm ends when sufficient training is in progress (the error converges to 1% or less between the algorithm output value Y1: YN and the TCAD simulation result value G1: GN). Therefore, the output value of well-training algorithm form a characteristic distribution similar to the TCAD simulation data [1, 7].

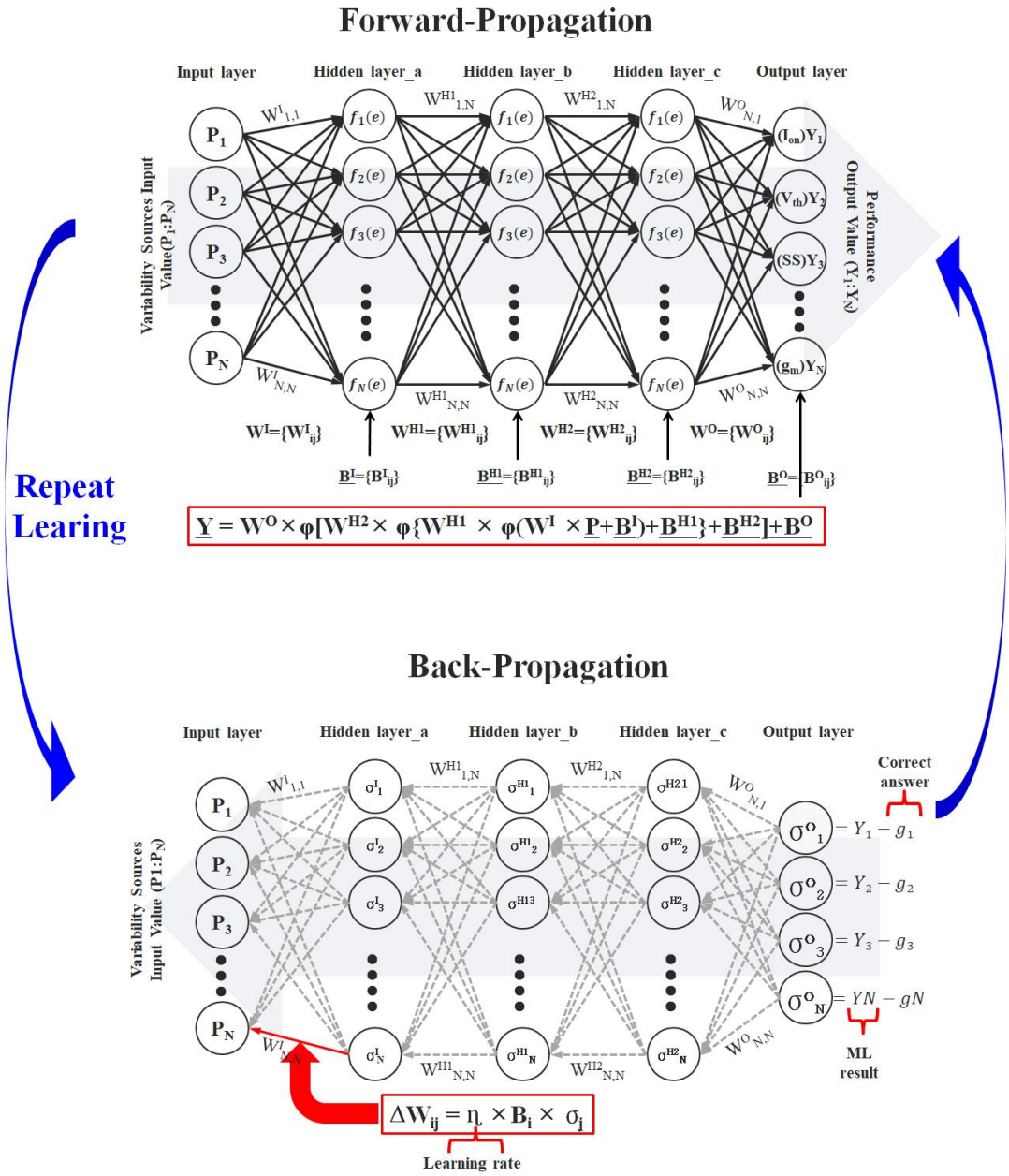


Fig. 2.6. Training schematic of artificial neural network algorithm.

The input and output data used in the ANN algorithm and the TCAD simulation can be described sequentially through the following algorithm application steps:

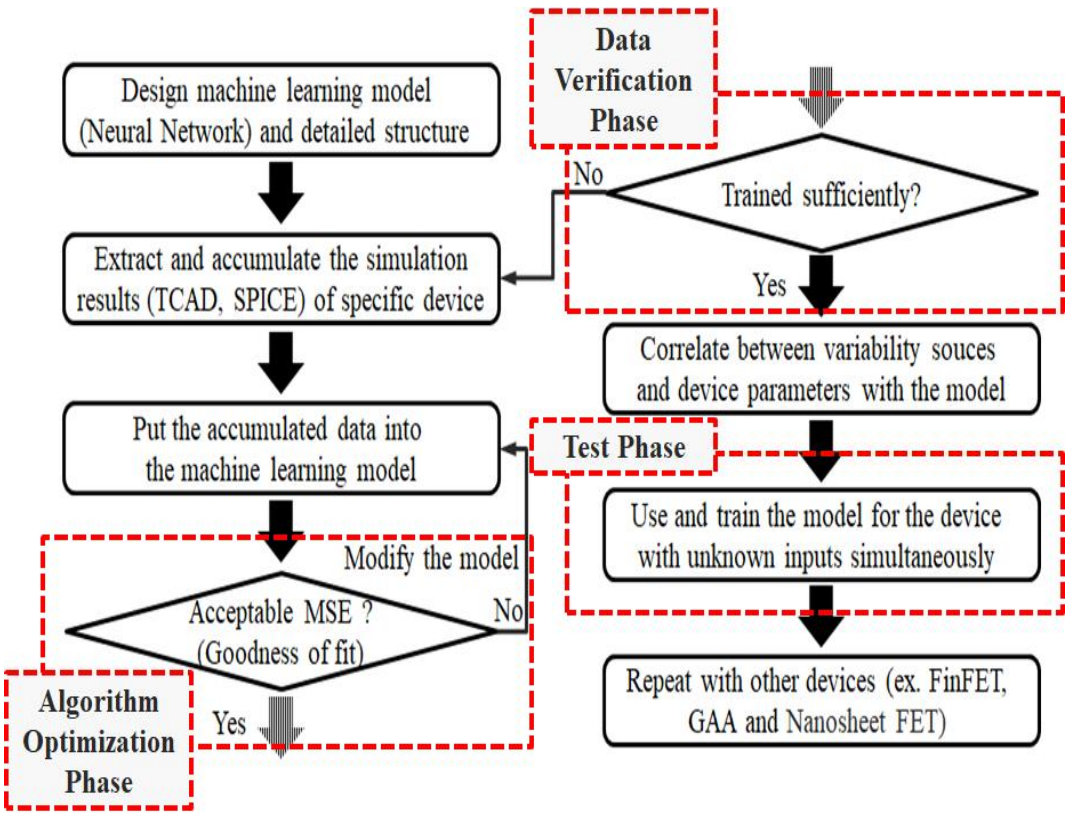


Fig. 2.7. The machine learning algorithm training procedures.

First, set up the basic algorithm structure and internal functions of the ANN algorithm. The training input/output data is extracted by TCAD simulation for training the ANN algorithm. In the training data extraction phase, the model trains the same value that were

input to the TCAD simulation. Therefore, the trained ANN algorithm adjusts the internal model based on the input and output data obtained from the TCAD simulation. In the optimization phase of the following algorithm, the output values of the ANN algorithm and the TCAD simulation are compared to adjust the weight of the internal function. Once the training algorithm is sufficiently completed (TCAD simulation data and error rate < 1%), the ANN algorithm is validated through the independently extracted test data. As for the verification result, the accuracy is determined by extracting the characteristic change data (output data for testing) corresponding to the implementation data (input data for testing) of the variability sources. In particular, in the verification process with test data, independent data is used rather than input/output data used in training. Therefore, it inserts new variable input data that has not been previously validated (not extracted by TCAD simulation) into the TCAD simulation and ANN algorithms simultaneously. After comparing the output values of the TCAD simulation and the ANN algorithm, the error rates of the two approaches are determined [8, 9].

The training and testing procedure of the ANN algorithm are performed as described above, and the input and output data used at each step are extracted and verified through the TCAD simulation. Therefore, the ANN algorithm training and testing method based on TCAD simulation should be considered as follows:

1) Quantitative preprocessing of input and output data.

Preprocessing of the data is required when the data used to train and test the ANN algorithm is extracted by TCAD simulation. TCAD simulation data that has not been

preprocessed is not suitable data type to enter the ANN algorithm. For TCAD simulations that use the 3D Stochastic Model to implement variability sources, the direct input of variability parameters is restricted. For example, to set the effect of WFV on TiN metal materials, we can perform TCAD simulations by providing arbitrary parameters (e.g., metal crystal generation seed, average metal crystal size, crystal distribution standard deviation, etc.) through an internal model. Therefore, <111> orientation and the <200> orientation metal grains are generated on the surface of the metal and the oxide. However, the parameters provided through TCAD simulation are not suitable as input/output data sets for training ANN algorithms. This is because data for training and testing ANN algorithms must consist of input and output data matrices of the same dimension. For example, if 1000 threshold voltage distribution is the output values, we need 1000 input values. Therefore, the input values of the variability source used in this study are the internal functions provided by the TCAD simulation, as well as the method of declaring arbitrary variables created in Python as external variables in the TCAD simulation.

2) Accuracy of TCAD simulation.

The comparison of ANN algorithms proposed in this paper is the predictive data of the effects of variability sources through TCAD simulation. Therefore, before applying the ANN algorithm, it is important that the analysis of the effects of the variability sources through the TCAD simulation is consistent with measurement results. Unfortunately, the results of all the variability sources presented in this paper are based on input and output

values through TCAD simulation and do not include actual measured data. However, based on the results of previous studies, the effect of variability sources through TCAD simulations are expected to be similar to actual device measurement results. Sections 3 and 4 of this study describe in more detail how to calibrate TCAD simulations and measurement data.

As a result, in this study, we optimize the ML approach to efficiently predict/analyze the various conditions above. Therefore, we have proposed an ML model that shows the same accuracy and fast predictability as the TCAD simulation.

2.4 Summary

To design an ML system that interoperates with TCAD simulation, it is important to develop detailed design and mapping technology inside the system to efficiently use the extracted input/output (I/O) training data. Therefore, in this paper, we focuses on the following four factors to develop the mapping technology of ML system.

1. Optimize the number of I/O data. Since I / O data is extracted using TCAD-based simulation and is directly related to the efficiency of the entire system, it is very important to find the optimal number of data considering the calculation cost.

2. Optimization of the number and number of hidden layers. For artificial neural network (ANN) -based algorithms, the number of computational nodes and the number of layers present in them are important factors in determining the complexity and accuracy of the overall algorithm. So the goal is to develop the optimal detailed design to reach the target accuracy.

3. Selection of forward calculation model and backward correction model. The basic operation mechanism of Artificial Neural Network (ANN), requires the selection of a forward computational model and a backward correction model as internal computational functions in order to represent an accurate predicted output value at an input value. The contents of internal functions differ depending on the purpose of use, so it is important to set clear goals for analysis and set the most efficient model.

4. Optimization of the number of training. For ML system, the more accurate the

training is, the more accurate the system. However, excessive iterative training can increase the computational cost of the system and can cause problems like vibration of the predicted output value, so the goal is to set the optimal number of iterations.

As a result, this study is divided into [Identify-Analysis-Interlock] step for extracting training data by TCAD simulation. We also propose an ML system based on the variability source analysis system that consumes optimal computational costs and accommodates the accuracy and integration of commercial tools.

References

[1] V. Subramanian, Deep Learning with Pytorch, 1st ed. iG Publishing Pte. Ltd. Press, 2018.

[2] ITRS, Denver, CO, USA. International Technology Roadmap for Semiconductors(ITRS), 2015. Online(http://www.itrs2.net/).

[3] Sentaurus Device User Guide Version: K_2015.06, Synopsys, Mountatin View, CA, USA, Jun. 2015.

[4] K. Ko, J. K. Lee, M. Kang, J. Jeon, and H. Shin, "Prediction of Process Variation Effect for Ultrascaled GAA Vertical FET Devices Using a Machine Learning Approach" IEEE Trans. Electron Devices, Vol. 66, No.10, pp. 4474-4477 Sep. 2019, doi: 10.1109/TED.2019.2937786

[5] V. Janakiraman, A. Bharadwaj, and V. Visvanathan, "Voltage and Temperature Aware Statistical Leakage," IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, Vol. 29, No. 7, pp. 1056-1069, Jul. 2010, doi: 10.1109/TCAD.2010.2049059

[6] J. Virarahgavan, S. J. Pandharpure, and J. Watts, "Statistical Compact Model Extraction:," IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, Vol. 31, No. 12, pp. 1920-1924, Dec. 2012, doi: 10.1109/TCAD.2012.2207955

[7] G. W. Burr, R. M. Shelby, S. Sidler, C. D. Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses),"

IEEE Trans. Electron Devices, Vol. 62, No.11, pp. 3498-3507, Nov. 2015, doi: 10.1109/TED.2015.2439635

[8] M. Abe, T. Nakamura, and K. Takeuchi, “Pre-shipment Data-retention/ Read-disturb Lifetime Prediction & Aftermarket Cell Error,” in Proc. Symp. VLSI Technol., pp. T216-T217, Jun. 2019, doi: 10.23919/VLS IT.2019.8776480.

[9] T. Ohashi, A. Yamaguchi, K. Hasumi, M. Ikota, G. Lorusso, C. L. Tan, G. V. D. Bosch, and A. Furnemont, “Precise measurement of thin-film thickness in 3D-NAND device with CD-SEM” Journal of Micro/Nanolithography MEMS and MOEMS, Vol. 17, No. 2, May. 2019, doi: 10.1117/1.JMM.17.2.024002

Chapter 3

Prediction of Process Variation Effect for Ultra-scaled GAA Vertical FET Devices

In this section, we present an accurate and efficient ML approach which predicts variations in key electrical parameters using PV from ultra-scaled GAA VFET devices. The 3D stochastic TCAD simulation is the most powerful tool for analyzing process variations, but for ultra-scaled devices, the computation cost is too high because this method requires a simultaneous analysis of various sources. The proposed machine-learning approach is a new method which predicts the effects of the variability sources of ultra-scaled devices. It also shows the same degree of accuracy as well as improved efficiency compared to a 3D stochastic TCAD simulation. An ANN-based ML algorithm can make MIMO predictions very effectively and uses an internal algorithm structure that is improved relative to existing techniques to capture the effects of process variations accurately. This algorithm incurs approximately 16% of the computation cost by predicting the effects of process variability sources with less than 1% error compared to a 3D stochastic TCAD simulation.

3.1 Introduction

Over the past four decades in the electronics industry, the typical size of an electronic device has been aggressively reduced to achieve higher levels of performance and efficiency, in keeping with Moore's law. Meanwhile, the aggressive trend of ultra-scaling the device size for integration is leading to unexpected problems, such as process variability issues. The major variability sources, including LV and GV sources, have been widely studied in relation to both planar and GAA devices [1-4]. In the case of LV sources, work-function variation (WFV) effects in high-k/metal gate (HK/MG) devices, which is noted as the most important statistical variability sources, should be considered to cope with a critical risk of device performance due to randomly occupied grain-orientation of gate metal [5, 8, 10]. Moreover, in the case of GV sources, the effect of CD variation is the most influential variability sources in varying the performance of the devices, and as the device size decreases, this effect can no longer be ignored. In earlier work, depending on the technology node, the structure of the device could change from the planar to the GAA type through the use of FinFET devices, and variability sources with dominant effects existed independently in each generation [6, 7]. However, in subsequent generations of transistors, the effects of all variability sources should be analyzed concurrently, as opposed to focusing on the key sources, for accurate predictions of the correlations for each source.

To analyze simultaneous variability in ultra-scaled devices, the 3D stochastic TCAD simulation has been proven to be superior in various studies as a leading analytical platform [3, 4]. However, as the number of variability sources to be considered increases, an analysis of a large number of sample devices is required. Therefore, a high computational cost is inevitable for accurate predictions. In recent studies, variability analysis methods with low computation overhead were proposed based on compact modeling for analysis of circuit level as well as device level, but overall this approach still has a high error rate because it is based on many preconditions for simplification of the formulae used [8-11]. Therefore, we propose a new variability analysis method based on a ML algorithm with accuracy that matches that of the 3D stochastic TCAD simulation in an ultra-scaled device and with improved calculation efficiency at the device level.

3.2 Simulation Structure and Methodology

The 3D TCAD simulation was performed using Synopsys Sentaurus, which was carefully calibrated with the experimental data of 5-nm 3stacked NP FET by referring to the reference [12]. To capture the transport characteristics of nanoscale devices, ballistic transport was considered, and the drift-diffusion (DD) approximation was used for carrier transport [13]. The saturated velocity and a ballistic model were carefully adjusted to match the experimental data. The interfacial mobility degradation was modeled using the Lombardi method, and a thin-layer mobility model was included to account for the thin channel thickness. In addition, the quantum effect caused by the small structure of the device is considered, and the contact resistivity, which has a large influence on the current as the device size is reduced, is set to $3 \times 10^{-9} \Omega/cm^2$ by referring to the reference [14].

Fig. 3.1 (b)-(c) shows the 3D structures of a NW VFET and a NP VFET, indicating the random distribution of metal grains on the HK/MG interface area, as noted in the TCAD simulation. The nominal device parameters are based on the International Technology Road Map for Semiconductors (ITRS) for 4/3 nm node technology low-power transistors, including a physical gate length (Lg) of 12 nm, an equivalent oxide thickness (EOT) of 0.6 nm, a nanowire diameter (dNW) and channel thickness (Tch) of 5 nm, channel width (Wch) of 40 nm, a supply voltage (Vdd) of 0.55 V.

Table 3.1 shows the quantitative values of the variability considered in this paper.

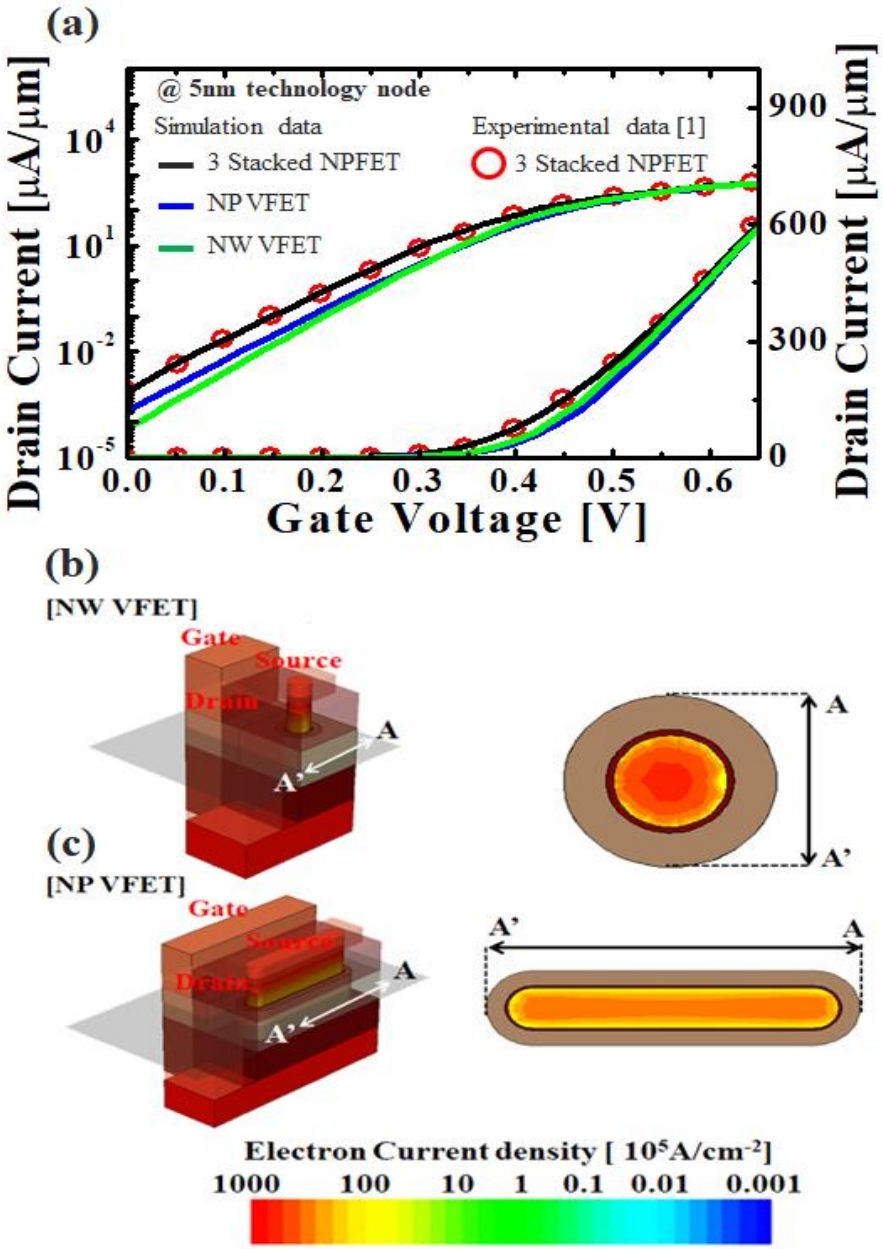


Fig. 3.1. (a) Calibration of the TCAD models with the experimental data of 5-nm node three stacked NP FET (b) Structure of NW VFET. (c) NP VFET.

In order to implement CD variation, the variation of Lg, Tch is assumed to be identical Gaussian distributions of 3σ=1 nm(critical dimension control). The variation of EOT is assumed to be 4% at 3σ [17]. In addition, the material of the metal gate is TiN, and the grain size (GS) set to 10 nm. The work function variation is implemented in random devices using Metal Work Function Variation (3D Stochastic WFV Model). Then, the electrical characteristics (Id-Vg plot) of the device were directly sampled through the TCAD Sentaurus workbench to extract the key electrical parameters caused by the WFV and CD variations. The key electrical parameters of the random device are extracted from approximately 1000 Id-Vg plot samples and are based on data that implements the individual dimensions of each device.

Process Parameters		
Critical Dimension Variations		
	Standard Deviation ($3\sigma^p$)	Average (μ^p)
L_g [nm]	1	12
T_{ch} [nm]	1	5
EOT [nm]	0.04	0.6
WF [eV]	60	4.6

Table. 3.1. Summary of the process variability sources parameters.

3.3 Results and Discussion

Figure 3.2 shows the basic structure of a MIMO ANN. The ANN algorithm structure has five layers, referred to as the input, first hidden, second hidden, third hidden, and the output layers. Each hidden layer that receives input layer data returns a calculated value to the output layer. For accurate process variation training, a rectified linear unit (RELU) activation function is used to prevent the vanishing gradient problem, which can arise due to the complexity of the algorithm. The ANN-based ML algorithm was trained using the PyTorch python library.

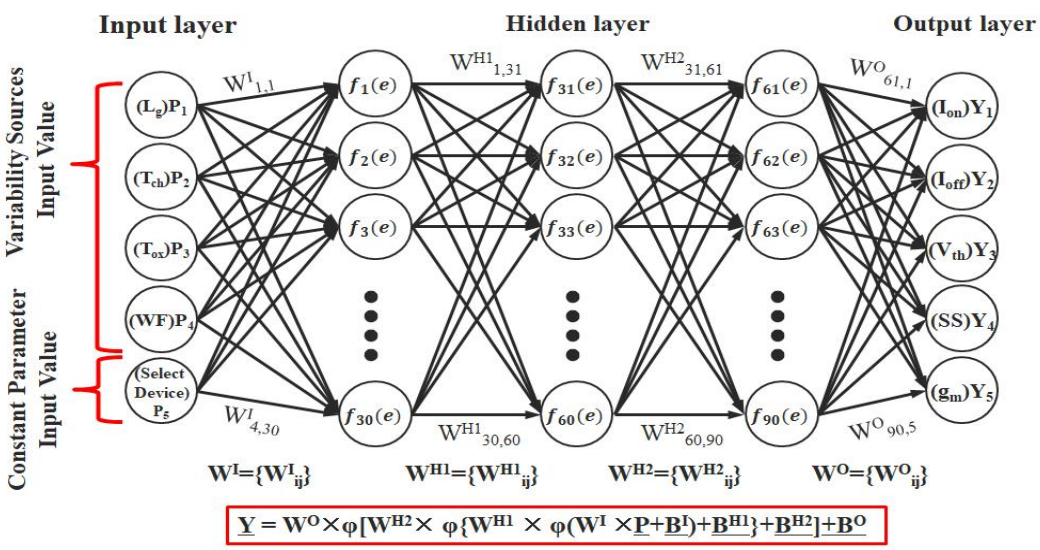


Fig. 3.2. MIMO artificial neural network structure. All input/output (I/O) values include the source of the process variation and the dimensions of the device and represent corresponding key electrical parameters.

Algorithm Building a MIMO ANN model

1: Generate minimum training samples from the distribution of process variation(I/O)
2: Generate 1000 testing samples from the distribution of process variation(I/O)
3: Initialize ANN parameters
4: **for** Number of training iterations **do**
  for $i = 1,2...n$ **do**
    $Y = W^{O} \times \varphi[W^{H2} \times \varphi\{W^{H1} \times \varphi(W^{I} \times \underline{P} + \underline{B^{I}}) + \underline{B^{H1}}\} + \underline{B^{H2}}] + \underline{B^{O}}$
5: Extract ANN output data from test samples
6: **if** ANN error rate(MSE) $\leq$ 1% **then**
  Done
 else Increase training iterations and go to Step 4
7: **else if**

Table. 3.2. Multi-Input Multi-Output ANN Algorithm.

1) W^{I} is 5 × 30 input weight matrix and B^{I} is a 30× 1 bias vector

2) $W^{H1,2}$ is 30 × 30 input weight matrix and $B^{H1,2}$ is a 30× 1 bias vector

3) W^{O} is a 5×30 output weight matrix and B^{O} is a 5×1 bias vector

4) (W^{I}, B^{I}), ($W^{H1,2}$, $B^{H1,2}$), (W^{O}, B^{O}) are unknown variables and are obtained during the trainging phase.

5) φ(x) is the activation function : Relu

6) P is a input vector (Variability parameters).

The weight value(W^{I}, W^{H1}, W^{H2}, W^{O}) assigned to each computation node and the vector value(B^{I}, B^{H1}, B^{H2}, B^{O}) that determines the computation direction are unspecified

variables and are constantly modified in the training phase. Therefore, the input and output values are calculated as a simple function, as shown in Figure 3.2. This calculation process is repeated approximately 20,000 times or more and the calculation runs in a form similar to the Taylor series to return an accurate final value [15, 16].

The input values inserted in the ANN algorithm include the work function variation value due to the WFV effect and the various values of Lg, Tch, and EOT due to the CD Variation. In addition, it contains the constant parameters of the specified device to learn various dimensions of the device. Therefore, although the structure of the device is different through single learning, it is possible to confirm the influence of the variability source in the NP VFET structure as well as the NW VFET. In addition, for the independent algorithm learning environment, the algorithm learns arbitrary values that are not related to the device specification presented above. The algorithm presented in this paper refers to the learned values of the GAA VFET device in the 6/5nm technology node [17]. The weights and vectors within the learned algorithm are used to determine the variability of the NW VFET and NP VFET test devices.

Figure 3.3 shows the TCAD simulation results to be learned and tested in the ANN algorithm for 1000 NW VFET and NP VFET device samples. The effect of the variability sources is larger in NW VFETs with relatively small channel area. This tendency is trained through the various input parameters of the ML approach mentioned above.

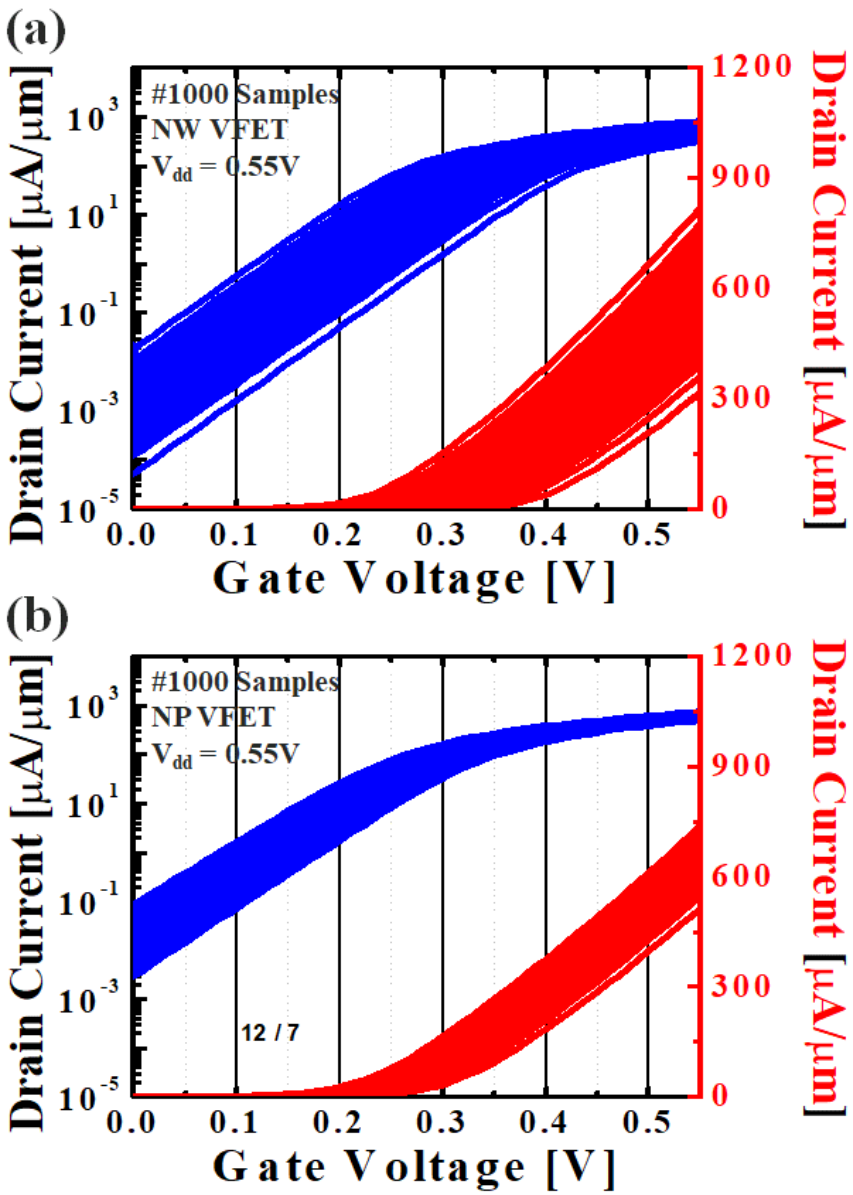


Fig. 3.3. Id-Vg plots for 1000 samples of (a) NW VFET and (b) NP VFET devices with all the variability sources.

Figure 3.4, 3.5 and 3.6 shows the distributions and the correlations between the key electrical parameter variations obtained from the 3D stochastic TCAD simulations and from the ANN-based ML algorithm. Both sets of distributions and correlations, including the effect of WFV and the effect of CD variations (i.e., ΔLg, ΔTch, and ΔEOT), achieve a good match between the TCAD data and the extracted ML data. In addition, the radar chart of the two approach shows that the rate of error is close to nil for all key electrical parameter variations. In the device specification of the same dimension, in the case of NP VFET structure, the influence of WFV is decreased due to the increase of the gate area, and the influence of CD variation is increased due to the decrease of the gate controllability, compared to the NW VFET. The change of the electrical characteristics by the WFV is a source that directly changes the threshold voltage of the device, so it does not affect the SS characteristic. Therefore, the correlations between SS and other performance characteristics are low for the NW VFET devices where the influence of WFV is dominant. However, for NP VFET devices with increased CD variation effects, there is a significant correlation between SS and other performance characteristics. As a result, the effect of the variability sources can be different depending on the structural characteristics of the device, and it can be confirmed that these differences are accurately predicted in the analysis through the machine learning algorithm.

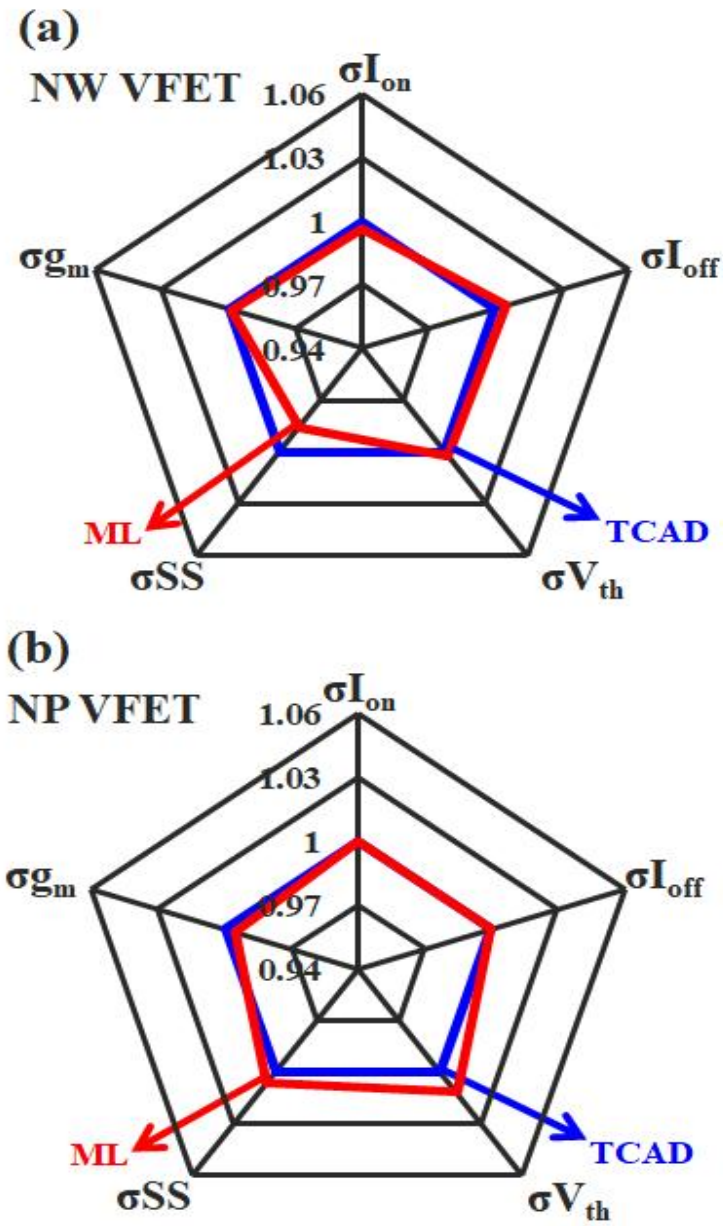


Fig. 3.4. Radar chart of the variation in key electrical parameters caused by WFV and GV sources (normalized by TCAD variations) of (a) NW VFET and (b) NP VFET.

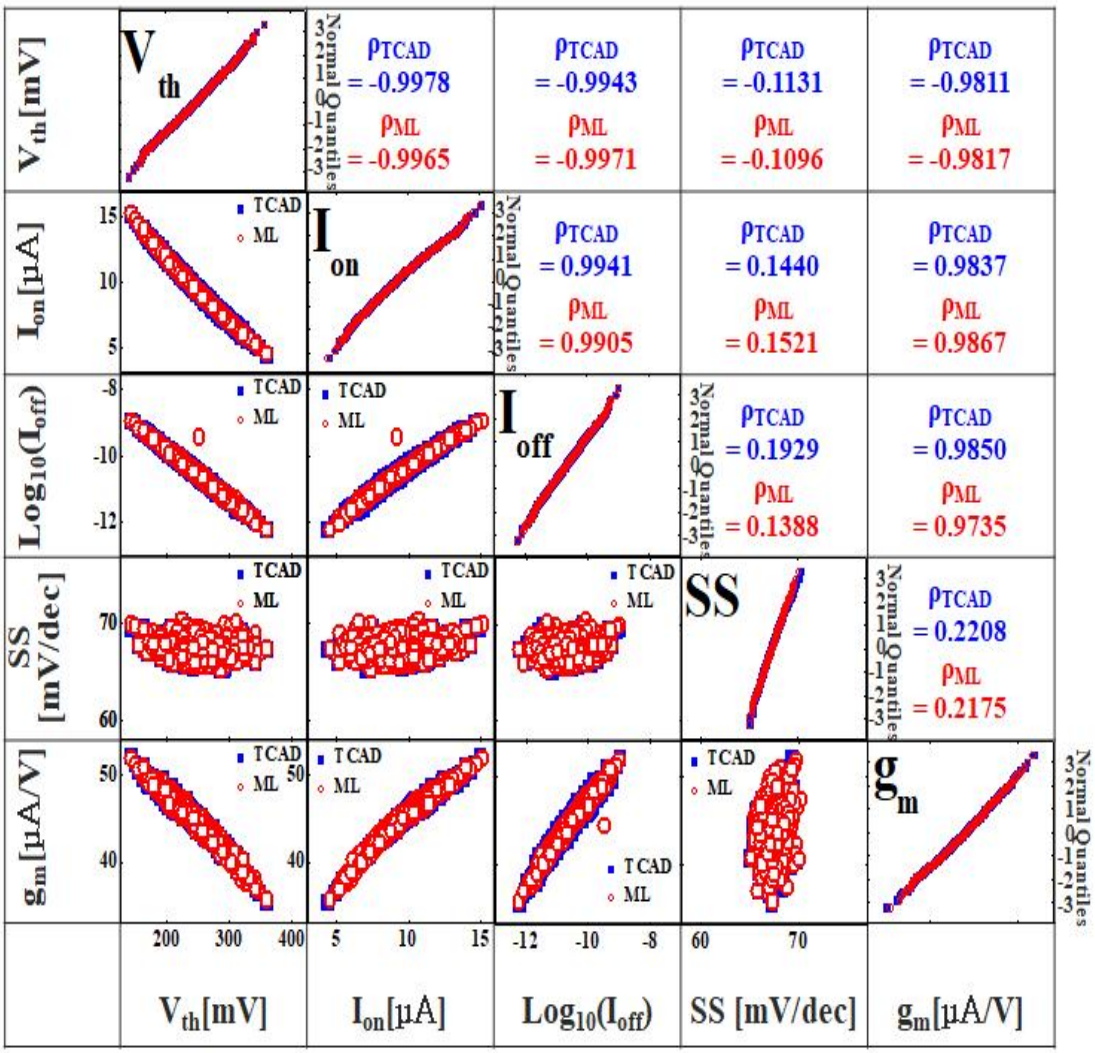


Fig. 3.5. Comparison of scatter plots and correlations of the key electrical parameters between the TCAD data and the ML Approach results from NW VFET. The Quantile-Quantile plots of the key electrical parameters are plotted in the diagonal subfigures.

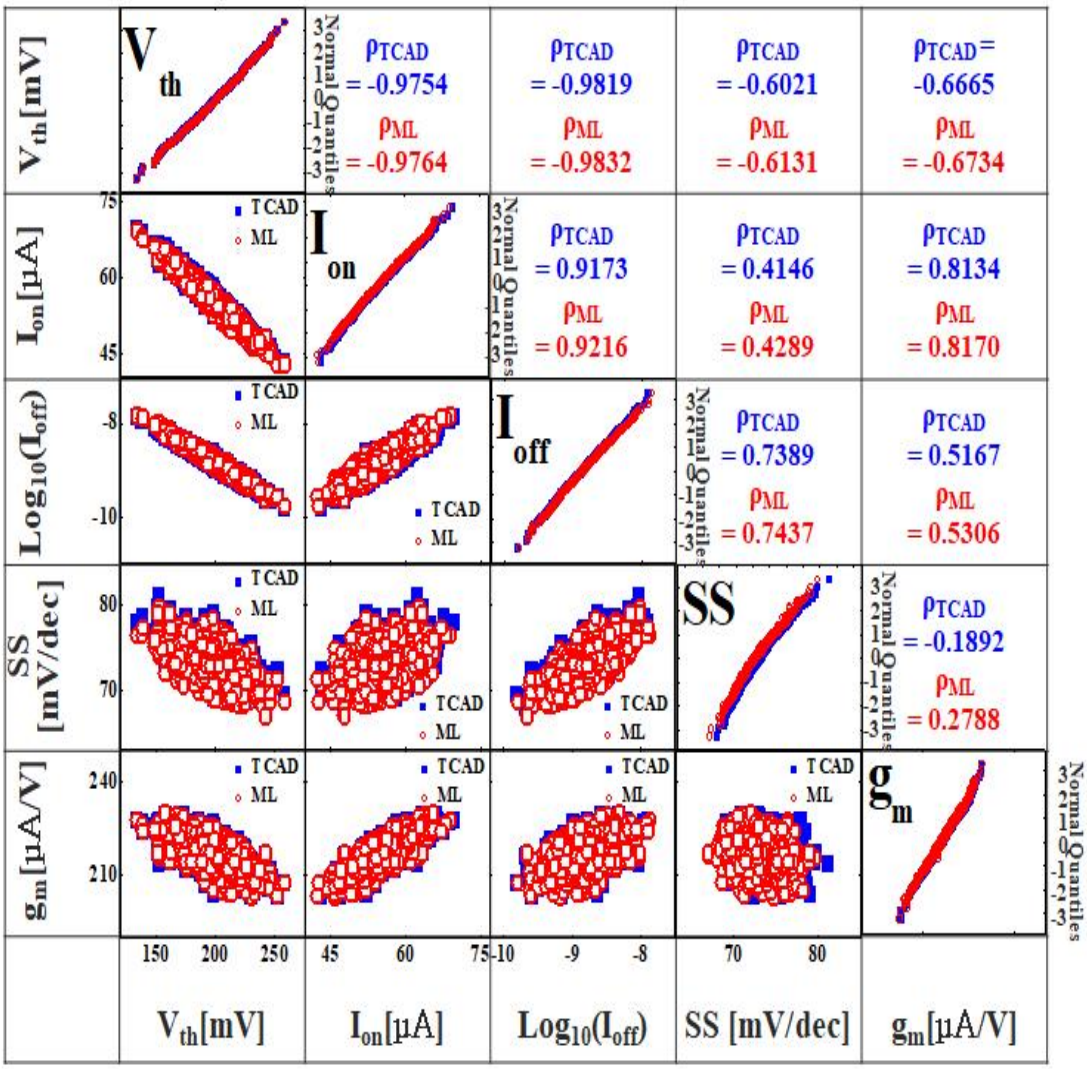


Fig. 3.6. Comparison of scatter plots and correlations of the key electrical parameters between the TCAD data and the ML Approach results from NP VFET. The Quantile-Quantile plots of the key electrical parameters are plotted in the diagonal subfigures.

The steps below are used for the quantitative extraction of the computation cost. The total computational cost of the ML approach for the analysis of process variability is determined by the sum of the amount of I/O data acquired for the training and the training and testing times of the internal algorithm. Therefore, in order to calculate the quantitative calculation cost compared to that of the TCAD simulation, the following three steps can be considered:

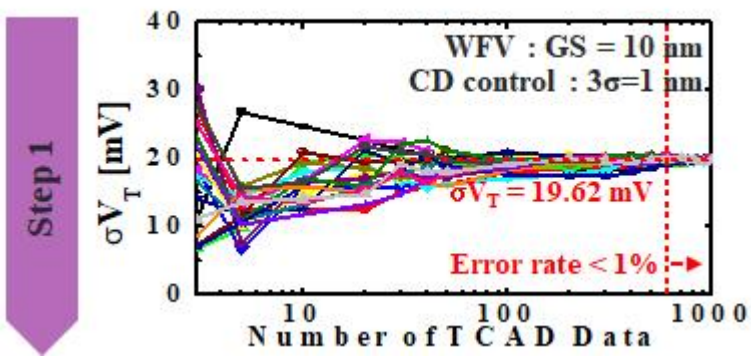


Fig. 3.7. Extract the minimum data amount to extract the effective distribution of the electrical characteristics

In the first step, the number of random devices samples required statistically is determined when analyzing the TCAD simulation independently. The variability sources used in this paper is the Gaussian distribution, which arbitrarily transforms the physical parameters of the device. Therefore, in order to understand the change in the characteristics of the device caused by such variability sources, a sufficient number of

random devices are required to form such a characteristic change as a Gaussian distribution. As a result, as shown in Step 1 of Fig. 3.7, when analyzing the characteristic change in about 600 or more random devices, it can be seen that the standard deviation converges to a certain Gaussian distribution. In other words, to analyze statistically the variability source, at least 600 TCAD simulation device analysis should be preceded. In this process, we can calculate the computational cost for analyzing the impact of variability using TCAD simulation.

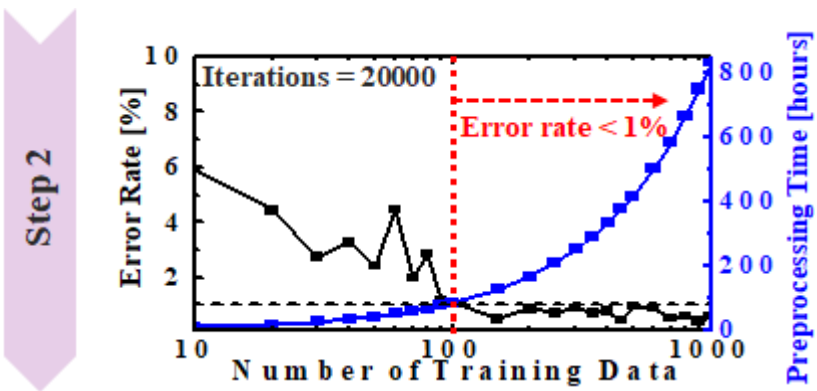


Fig. 3.8. Extract the training data amount of ML Approach to ensure the same level of accuracy as the TCAD simulation.

In the second step, the minimum number of random device samples for learning ANN algorithm through TCAD simulation data is grasped. In the first phase, a TCAD simulation required samples of at least 600 random devices to statistically analyze the variability. However, the ANN algorithm shows that sufficient prediction is possible with

fewer random device samples than the TCAD simulation. The number of minimum random device samples for learning the ANN algorithm presented in this paper is about 100. In this process, a random device sample of the TCAD simulation for training and testing exists independently. The number of random device samples for training is gradually increased. After the training, the ANN algorithm model extracts the error rate by comparing it with about 1000 random sample samples for the test. As a result, the analysis of the influence of the variability sources through the ANN algorithm requires about 6 times fewer random device samples than the TCAD method. Therefore, it is possible to quantitatively estimate the reduction in the computational cost of the ANN algorithm when the error of the two approaches is less than 1%.

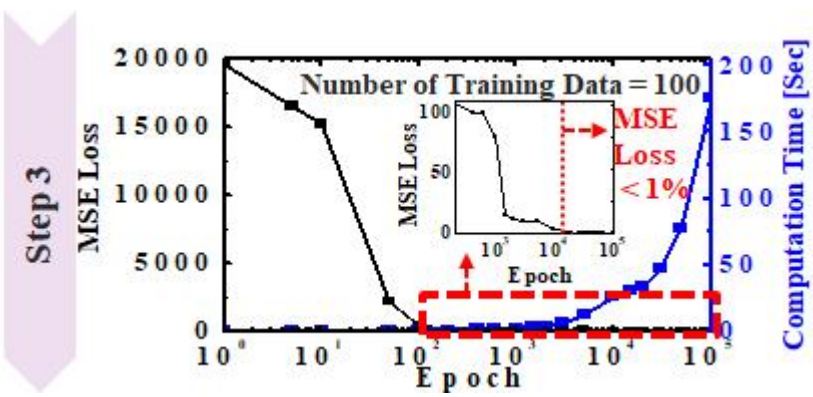


Fig. 3.9. Training and calculation time extraction of ML Approach internal algorithm based on acquired training data.

In the third step, the number of learning iterations for training 100 ANNs of random

device samples claimed in the previous step is shown. In the following procedure, we show the error with testing data (1000 random device samples for testing) according to the number of learning iterations of the ANN algorithm learned through training data (100 random device samples for learning). In the same way as above, the training data and the testing data exist independently and are extracted through TCAD simulation. As a result, the number of training iterations shows an effective reduction of the error rate when it is performed about 20,000 times or more.

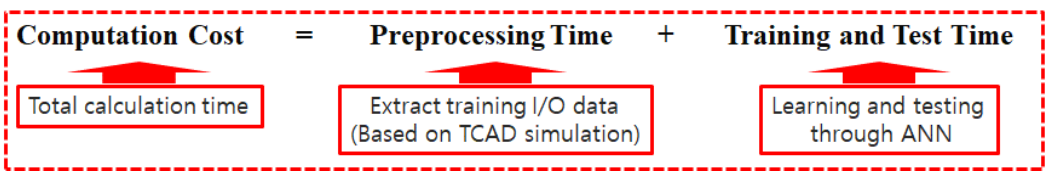


Fig. 3.10. Schematic diagram of total calculation cost.

Therefore, considering the total computational cost, the ML approach is nearly six times faster than the TCAD simulation, as shown in figure 3.10. The error rate of the key electrical parameter variation is trained to be less than 1% on average, as shown in Table 3.3.

Approach	Computation time (hours)	RMS error in σFoMs (percent error with mean)		
Stochastic 3D TCAD simulations	500	0 (Benchmark)		
This work	83.33	NW VFET	I_{on} [nA]	1.1004(0.06%)
			$Log_{10}(I_{off})$	0.0012(0.23%)
			V_{th} [μV]	2.378(0.01%)
			SS [μV/dec]	21.3664(2.59%)
			g_m [nA/V]	3.3678(0.13%)
		NP VFET	I_{on} [nA]	1.2471(0.02%)
			$Log_{10}(I_{off})$	0.0037(1.21%)
			V_{th} [μV]	0.6664(0.01%)
			SS [μV/dec]	13.3908(0.67%)
			g_m [nA/V]	21.0514(0.44%)

Table 3.3. Comparison of the computation cost and accuracy of ML approach and TCAD Simulation.

Compute Environment: Intel Xeon CPU E5-2687W 216 V2 (3.5 GHz × 16) processor with 128 GB of RAM

Stochastic approach (TCAD Simulation) takes about 0.83 h per single sample. Thus, analyzing 600 samples takes about 500 h. ML approach (MINO ANN Algorithm), it takes about 83.33 h to obtain minimum training data and calculate test data. In the ML approach compared to TCAD, the computational cost can be reduced by about 83%.

3.4 Summary

In this section, we proposed a new analytical method based on an ANN-based ML algorithm to assess the degree of process variability in ultra-scaled devices. According to previous studies, 3D stochastic TCAD simulations are considered to be the most powerful means of analyzing process variation. However, for an analysis of ultra-scaled devices, various process variation sources must be analyzed simultaneously and integrally. Therefore, existing methods require too expensive computation cost for accurate prediction. Therefore, analyzing the process variation sources using the ANN-based ML approach can significantly improve (6×) the calculation efficiency while maintaining the high accuracy (relative error of approximately 1%) of the existing 3D stochastic TCAD simulation. As a result, the method of analyzing process variability levels proposed in this paper can provide useful guidelines to those who develop and design ultra-scaled devices.

References

[1] X. Wang et al, “FinFET Centric Variability-Aware Compact Model Extraction” IEEE Trans. Electron Devices, vol. 62, no. 10, pp. 3139-3146, Oct. 2015, doi: 10.1109/TED.2015.2463073

[2] F. A. Lema et al, “Comprehensive ’Atomistic’ Simulation of Statistical Variability, and Reliability in 14 nm Generation FinFETs” in proc. IEEE SISPAD, pp. 157-160, Sep. 2015, doi: 10.1109/SISPAD.2015.7292283

[3] X. Wang, B. Cheng, A. R. Brown, C. Millar, J. B. Kuang, S. Nassif, and A. Asenov, “Interplay Between Process-Induced and Statistical,” IEEE Trans. Electron Devices, vol. 60, no. 8, pp. 2485-2492, Aug. 2013, doi: 0.1109/TED.2013.2267745

[4] L. Gerrer et al, “Interplay Between Statistical Reliability and Variability, ” in Proc. IRPS, Jun. 2013, pp. 3A.2.1-3A.2.5, doi: 10.1109/IRPS.2013.6531972

[5] X. Wang, A. R. Brown, B. Cheng, and A. Asenov, “Statistical variability and reliability in nanoscale FinFETs” in Proc. IEEE Int. Electron Devices Meeting, Dec. 2011, pp. 5.4.1-5.4.4, doi: 10.1109/IEDM.2011.6131494

[6] T. H. Bao et al, “A Comprehensive Benchmark and Optimization of 5-nm Lateral and Vertical GAA 6T-SRAMs” IEEE Trans. Electron Devices, vol. 63, no. 2, pp. 643-652, Feb. 2016, doi: 10.1109/TED.2015.2504729

[7] V. Moroz, “Transistor and Logic Design for 5nm Technology Node” Synopsys Silicon to Software, Semicon Taiwan , Sep. 2016

[8] H. Dadgour, K. Endo, V. K. De, and K. Banerjee “Grain-Orientation Induced Work Function Variation in Nanoscale Metal-Gate Transistors—Part I,” IEEE Trans. Electron Devices, vol. 57, no. 10, pp. 2504-2514 Oct. 2010, doi: 10.1109/TED.2010.2063191

[9] P. H. Vardhan, S. Mittal, S. Ganguly, and U. Ganguly, “Analytical Estimation of Threshold Voltage,” IEEE Trans. Electron Devices, vol. 64, no.8, pp. 3071-3077 Aug. 2017, doi: 10.1109/TED.2017.2712763

[10] K. Ko, M. Kang, J. Jeon, and H. Shin, “Compact Model Strategy of Metal-Gate,” IEEE Trans. Electron Devices, vol. 66, no.3, pp. 1613-1617 Mar. 2019, doi: 10.1109/TED.2019.2891677

[11] X. Jiang, X. Wang, R. Wang, B. Cheng, A. Asenov, and R. Huang, “Predictive compact modeling of random variations in FinFET technology for 16/14nm node and beyond” ” in Proc. IEEE Int. Electron Devices Meeting, Dec. 2016, pp. 28.3.1-28.3.4, doi: 10.1109/IEDM.2015.7409787

[12] N. Loubet et al., “Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET,” in Proc. Symp. VLSI Technol., Jun. 2017, pp. T230–T231, doi: 10.23919/VLSIT.2017.7998183.

[13] J. D. Bude, “MOSFET modeling into the ballistic regime,” in Proc. IEEE SISPAD, Sep. 2000, pp. 23–26, doi: 10.1109/SISPAD.2000.871197.

[14] P. Adusumilli et al., “Ti and NiPt/Ti liner silicide contacts for advanced technologies,” in Proc. IEEE Symp. VLSI Technol., Jun. 2016, pp. 1–2.

[15] J. Virarahgavan, S. J. Pandharpure, and J. Watts, “Statistical Compact Model

Extraction:," IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, Vol. 31, No. 12, pp. 1920-1924, Dec. 2012, doi: 10.1109/TCAD.2012.2207955

[16] V. Subramanian, Deep Learning with Pytorch, 1st ed. iG Publishing Pte. Ltd. Press, 2018.

[17] ITRS, Denver, CO, USA. International Technology Roadmap for Semiconductors(ITRS), 2015. Online(http://www.itrs2.net/).

Chapter 4

Prediction of Process Variation Effect for 3D NAND Flash Memories

In this section, we propose a variability-aware ML approach that predicts variations in the key electrical parameters of 3D NAND Flash memories. For the first time, we have verified the accuracy, efficiency, and generality of the predictive impact factor effects of ANN algorithm-based ML systems. ANN-based ML algorithms can be very effective in MIMO prediction. Therefore, changes in the key electrical characteristics of the device caused by various sources of variability are simultaneously and integrally predicted. This algorithm benchmarks 3D stochastic TCAD simulation, showing a prediction error rate of less than 1% as well as a calculation cost reduction of over 80%. In addition, the generality of the algorithm is confirmed by predicting the operating characteristics of the 3D NAND Flash memory with various structural conditions as the number of layers increases.

4.1 Introduction

In order to increase device integration, the Flash memory industry has evolved from planar to 3D-type NAND [2, 3]. The NAND flash memory devices are greatly integrated for process technology scaling and multi-level technology. The current technologies typically use poly-silicon as the material for the channel of the device [4]. Although using such a 3D structure and channel material is easy to integrate and cost-effective, there are various limitations regarding the material and 3D structure of the memory device [5]. These limitations can be accurately predicted by a simultaneous analysis of various variability sources, as opposed to simply considering the effects of several key sources, due to the continued scaling of the devices and ever-increasing number of layers. In a recent study, the variability of the 3D NAND flash memory devices was attributed to the following: the effects of Grain-Boundary Traps (GBT) caused by the poly-silicon channel, Critical Dimension (CD) control fluctuations due to process limitations, and tapered channel effects occurring with increasing numbers of layers [6-11]. Therefore, when analyzing these various sources simultaneously, the commonly used 3D TCAD simulation may suffer from an efficiency problem due to the excessive calculation cost. As mentioned above, in order to accurately predict the characteristics change of NAND flash memory devices according to the various variability sources, it is appropriate to analyze them through the distribution of large sample devices. In other words, as the variability

source to consider increases, more and more sample device data must be analyzed. Therefore, high computational costs are inevitable to ensure high prediction accuracy. Therefore, a new predicting and analysis system that can predict various variability sources more quickly and accurately is necessary. As a result, this paper presents a new ML approach that can predict rapidly and with the same accuracy as the 3D stochastic TCAD simulation.

4.2 Simulation Structure and Methodology

4.2.1 3D Device Structure and variability sources

A schematic and a cross-sectional view of the stacked 3D NAND flash memory used in 3D TCAD simulation is shown in figure 4.1. It consists of a string array of vertical channel charge trap (CT) devices with a gate stack of oxide/nitride/oxide (O/N/O) and a Macaroni body. Each string consists of metal plate word-lines (WLs) and a select transistor for each of the bit-line (BSL) and the source-line (SSL). Table 4.1 shows the device parameters for simulation. The channel cross-section presented in figure 4.1 directly shows the variability sources set in this paper. In total, eight variability sources are simultaneously considered, and each variability source refers to an externally declared input value through Python.

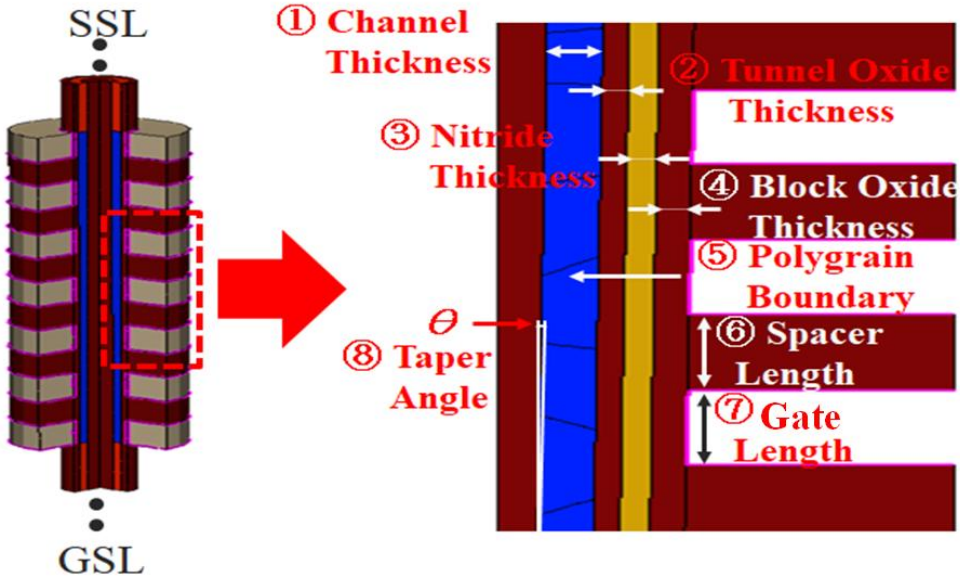


Fig. 4.1. Schematic of BiCS 3D NAND architecture and the tapered vertical NAND flash memory string and cross-section of the control gate part.

Device parameters	Value	Device parameters	Value
Filler Oxide Thickness (T_{filler}) (nm)	30	Channel doping (cm^{-3})	10^{16}
Channel Thickness (T_{ch}) (nm)	10	Grain Size (nm)	30
Gate Length (L_G) (nm)	28	Tunneling Oxide Thickness (nm)	5
Spacer Length (L_S) (nm)	28	Nitride Thickness (nm)	5
Metal Thickness (T_{metal}) (nm)	40	Blocking Oxide Thickness (nm)	6

Table. 4.1. Summary of device Parameters.

It is also assumed that each of the variability forms an ideal Gaussian distribution, and the standard deviation is set based on the International Technology Road Map for Semiconductors (ITRS) 2.0 2015 recommendation and reference [6-11]. The 3D TCAD simulation was performed using Synopsys Sentaurus, which was carefully calibrated with the experimental data regarding the 3D NAND string obtained from reference [1], as shown in figure 4.2. The simulated structure (figure 4.1 and figure 4.2) resembles the Macaroni body vertical channel CT memory obtained using punch-and-plug process utilized for fabricating 3D NAND flash memories. Although grain boundaries in the polysilicon channel are known to affect the string electrostatics and increase the device variability. Therefore, we implement the grain boundary in three dimensions by using external variable declaration and apply it to the simulation. Drift-diffusion based simulations were performed assuming a constant mobility and utilizing Shockley-Read-Hall (SRH) recombination model. The voltages of all 10 WLs were ramped together (multi-WL measurement) to explore the average behavior of the string current [1]. In addition, the mobility model is constructed by referring to the previous studies mentioned in reference [1]. A constant mobility μ_n = 100 cm^2 / Vs, falling in the range typically assumed in the literature [13], [16], was used for electron drift / diffusion inside the grains (no mobility reduction was introduced close to the grain boundaries [9]). Therefore, the physical model of TCAD simulation presented in this paper preferentially uses the model verified by comparison with measurement data in the previous study. It may be noted that the main purpose of this work is to propose a method for formulating a ML approach of

3D NAND flash rather than showing the exact values of the string current. Therefore, figure 4.3 and figure 4.4 shows that the impact of all of the variability sources is significant when analyzing the independent effects of each variability. This confirms that, in the case of a 3D NAND flash memory device, all of the variability sources must be taken into account, as opposed to considering only the dominant sources.

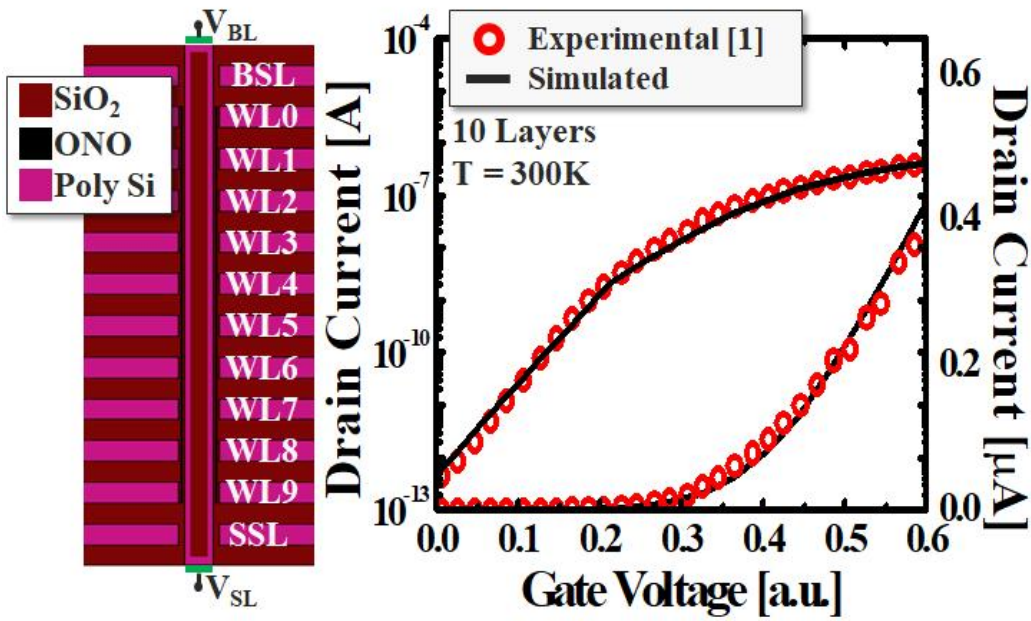


Fig. 4.2. Calibration of the TCAD Simulation set up by reproducing the experimental string-current multi-WL characteristics of 10 WL string.

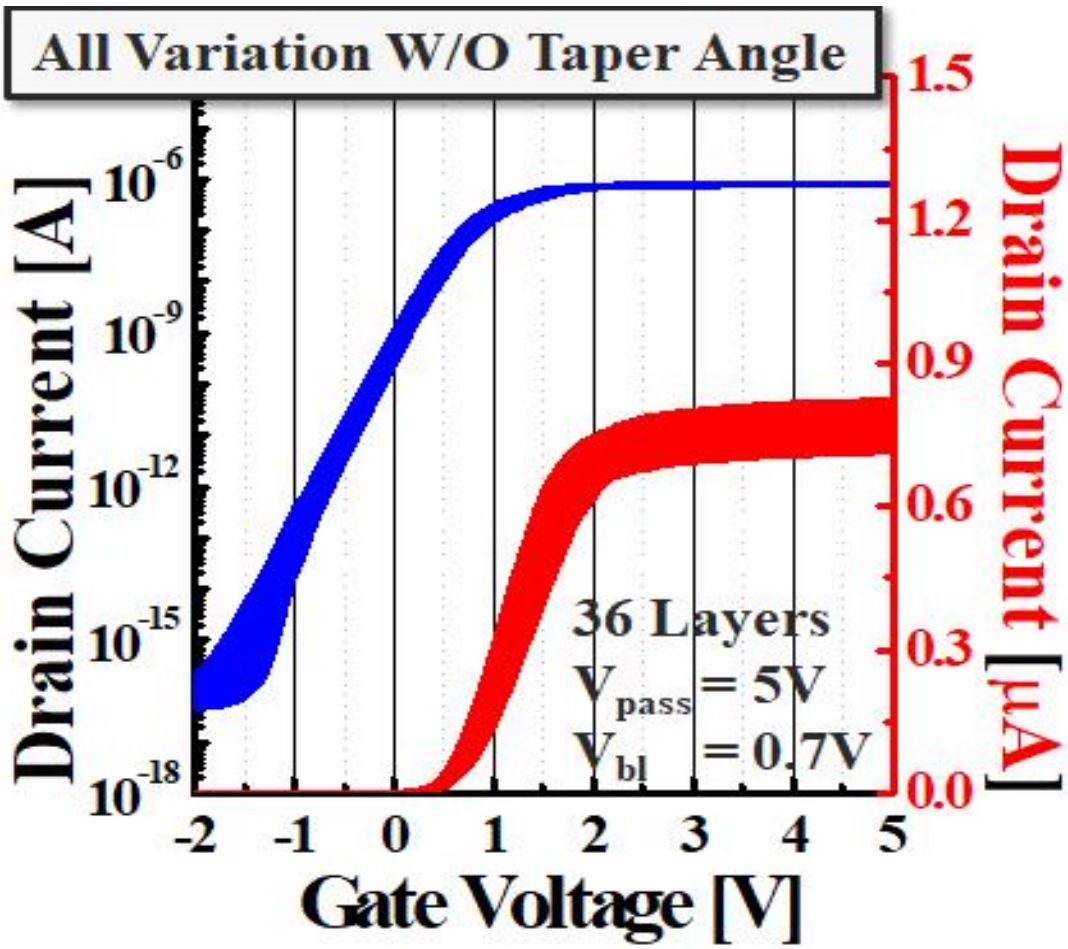


Fig. 4.3. Id-Vg plots for 600 samples of 3D NAND Flash memory devices with all the variability sources except the effect of taper angle.

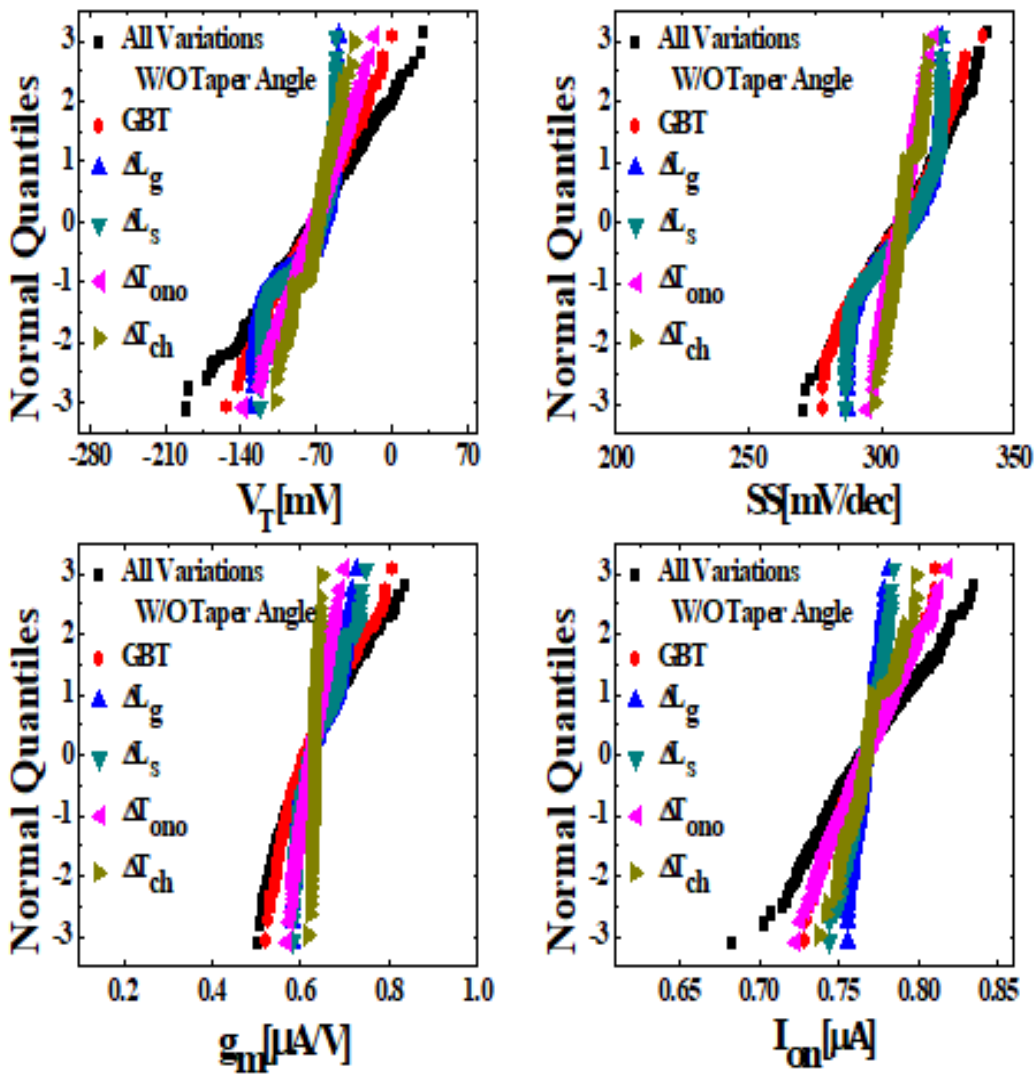


Fig. 4.4. The quantile-quantile plot (Q-Q plot) for the key electrical parameters (Vth, SS, gm, Ion) shift due to individual and combined effects of five variability sources.

4.2.2 Machine Learning Modeling Using ANNs

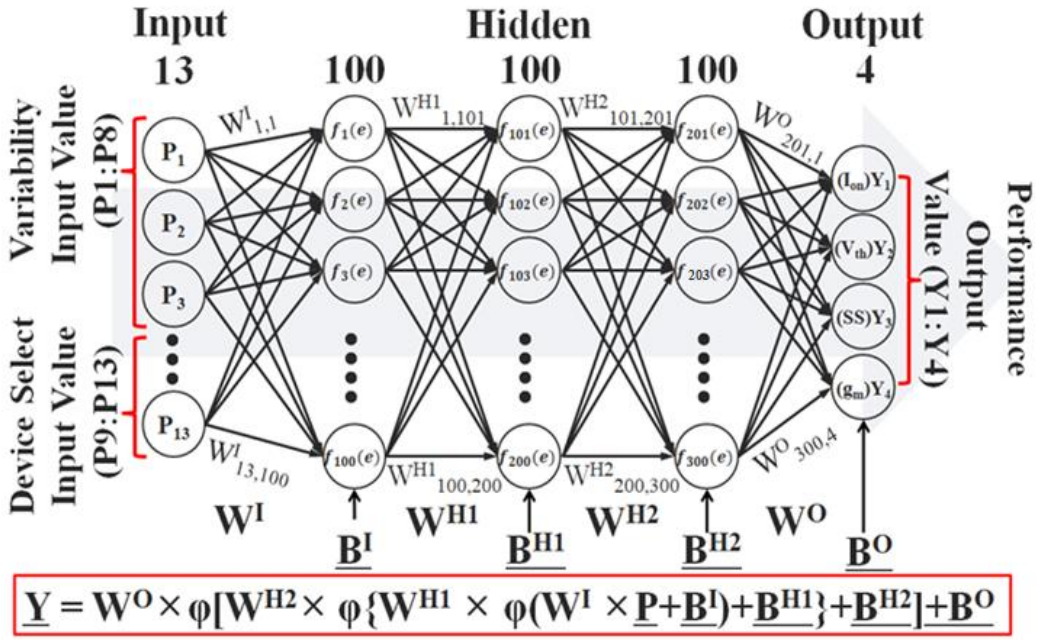


Fig. 4.5. MIMO artificial neural network structure. All input/output values include the source of the process variation and the dimensions of the device and represent corresponding key electrical parameters.

Figure 4.5 shows the basic structure of a MIMO ANN. As shown in Table 4.2, the ANN algorithm structure has five layers, which are referred to in order as the input, first hidden, second hidden, third hidden, and output layers [12-15]. In particular, the input data of the ANN algorithm presented in this paper can be divided into two types:

Algoritm parameters	Value	Algoritm parameters	Value
Device Select Parameters (N_{WL},T,V_{BL}, V_{WL}, Selected WL)	5	No. of inputs (P)	13
		No. of hidden layers (H)	3
Variability Sources Parameters (GBT, $T_{o\_tunneling}$, T_n, T_{o_block}, T_{ch}, L_g, L_s, Taper angle)	8	No. of hidden nodes (N)	300
		No. of outputs (Y)	4

Table. 4.2. Summary of the artificial neural network (ANN) training process parameters.

1) Variability sources input data and 2) Constant parameter input data. The variability sources input data are data used to correlate the direct variation of the variability sources with the change in device characteristics, and it includes the physical factor change of the device due to the direct variation source. On the other hand, the constant parameter input data are values for distinguishing the structure and electrical characteristics of the analyzing device, and it includes fixed parameters. By using these two types of input data, we can train the network to obtain the effect of the variability of 3D NAND flash memory devices on the ANN-based ML algorithm under various conditions. Each hidden layer that receives input layer data propagates a calculated value to the output layer. For accurate variation training, a Rectified Linear Unit (RELU) activation function was used to prevent

the vanishing gradient problem that can be caused by the complexity of the algorithm. The weight value (WI, WH1, WH2, WO) assigned to each computation node and the vector value (BI, BH1, BH2, BO) that determines the computation direction are unspecified variables that are constantly modified in the training phase [14, 16]. The Adaptive Moment Estimation (ADAM) optimization function was used for accurate error correction. The ADAM optimization function is an optimization function that is known to be suitable for processing stochastic data and that corrects the learning of the algorithm through iterative correction by the set error rate (η) [13, 16]. The ANN-based ML algorithm was trained using the PyTorch python library. Therefore, the input and output values are calculated as a simple function, as shown in Figure 4.5. This calculation process was repeated approximately 20,000 times or more, and the calculation runs in a form similar to the Taylor series to return an accurate output value [14-16].

Figure 4.6 shows the training process through the interlocking with the 3D TCAD simulation of the ML algorithm. At each step, the ANN-based ML algorithm shares data on the 3D TCAD simulations continuously and repeatedly for sufficient training. As a result, once it has completed training, the ML algorithm can independently predict and analyze the effect of the variability sources in various 3D NAND Flash memories.

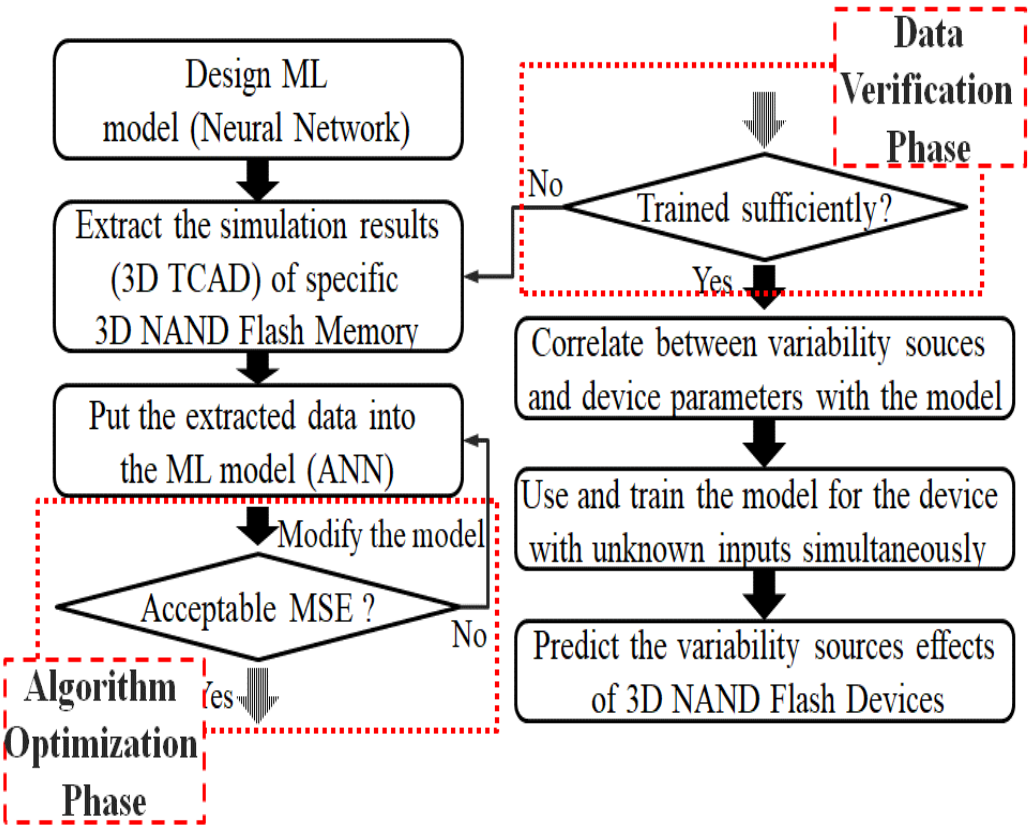


Fig. 4.6. The ML algorithm training procedure that benchmarks the 3D TCAD simulation is adjusted to the error of the result. 600 simulations are used for the configured device dimensions.

4.3 Results and Discussion

4.3.1 Statistical Analysis with ANNs

Table 4.3 summarizes the process variability sources parameters. The variability formed by the ideal Gaussian distribution includes seven variability sources aside from tapered channel effects [6-11]. The characteristic change due to the tapered channel effect is also an important factor that can have a critical effect on NAND Flash memory devices. However, too many cases have to be taken into account for analysis in all sources, so representative cases are selected for analysis and prediction. Therefore, the characteristics variation analysis and prediction accuracy of the device are extracted by first considering seven variability sources except the tapered angle effect. Each variability source has an independent effect on an arbitrary string, and its corresponding impact is analyzed using a large number of sample devices. Therefore, if the various variability sources independently affect the device, this can be confirmed through the distribution of key electrical characteristics such as Vth, SS, gm, and Ion.

Process Parameters			
Critical Dimension Variations			
	Standard Deviation (3σp)		Average (μp)
L_g [nm]	1.2		28
L_s [nm]	1.2		28
T_{ch} [nm]	1.2		10
$T_{tunneling\_oxide}$ [nm]	0.69		5
$T_{nitride}$ [nm]	0.69		5
$T_{blocking\_oxide}$ [nm]	0.69		6
Polysilicon Grains			
Grain Size [3σ,nm]	15	Average Grain size [μ,nm]	30
Grain Angle [3σ,rad]	1	Average Grain Number [#]	75

Table 4.3. Summary of the process variability sources parameters.

Figure 4.7 shows the distribution of the key electrical characteristics due to seven variability sources and shows that the distribution of characteristics through the ML algorithm is also accurately predicted. Compared with the 3D TCAD simulation results, 600 test data independent of the training data are used to predict the effect of the variability sources through the ML system. The amount of change is calculated based on the average of the electrical characteristic distribution of the NAND Flash memory device. Therefore, the distribution of four electrical characteristics such as Vth, Ion, Ioff, and gm is predicted as shown in the figure, and both ML system and TCAD simulation results show the distribution of electrical characteristics in the form of an ideal Gaussian distribution.

Comparing the predicted characteristic distribution through ML with the results extracted through 3D TCAD reveals that the average error rate of the standard deviation of the key electrical characteristic distribution is less than 1%, as shown in Figure 4.8. Independent training data and test data sets are considered to verify the error rate within 1%. First, TCAD simulations generate 100 training data sets and then form sufficiently internal models through iterative learning of ML systems. After that, we extract the results of the TCAD simulation using 600 independently generated test data sets. At the same time, the output of the ML system with 600 inputs of variability sources is extracted and compared with the TCAD simulation results. As a result, it can be seen that the error rate is formed within 1% for 600 test data sets.

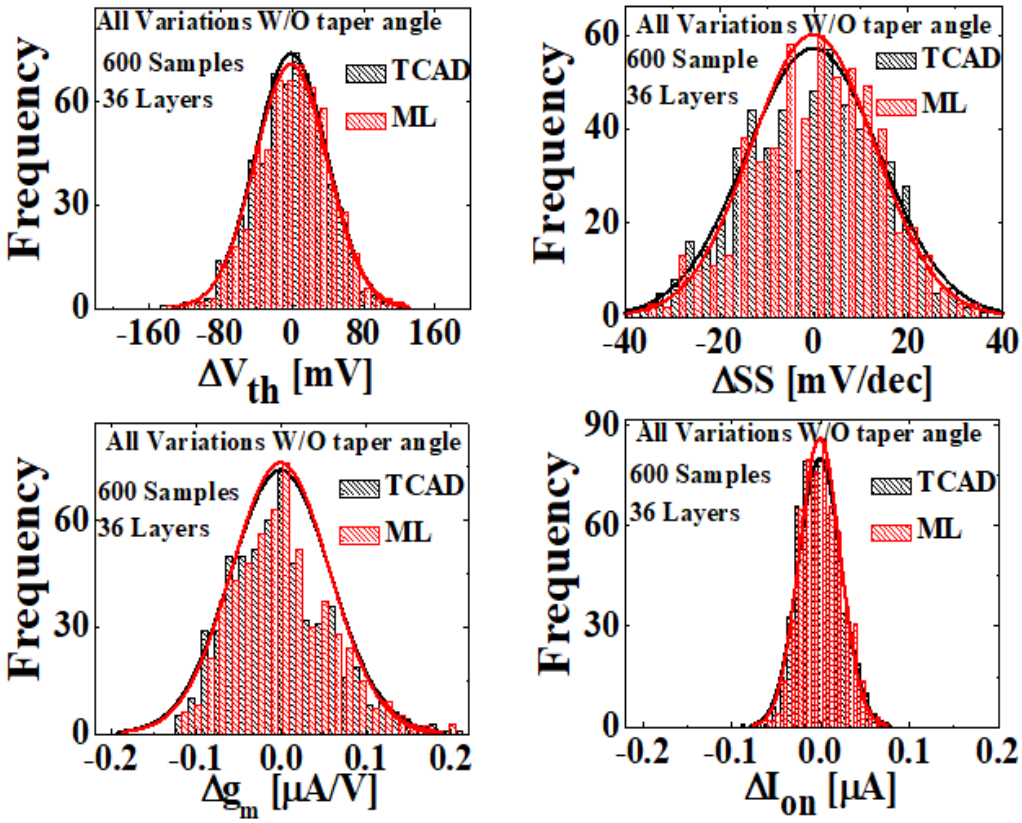


Fig. 4.7. The key electrical parameters distribution for 600 samples of 3D NAND Flash memory devices. All distributions are formed in normal distribution and show the correspondence between ML algorithm and TCAD simulation.

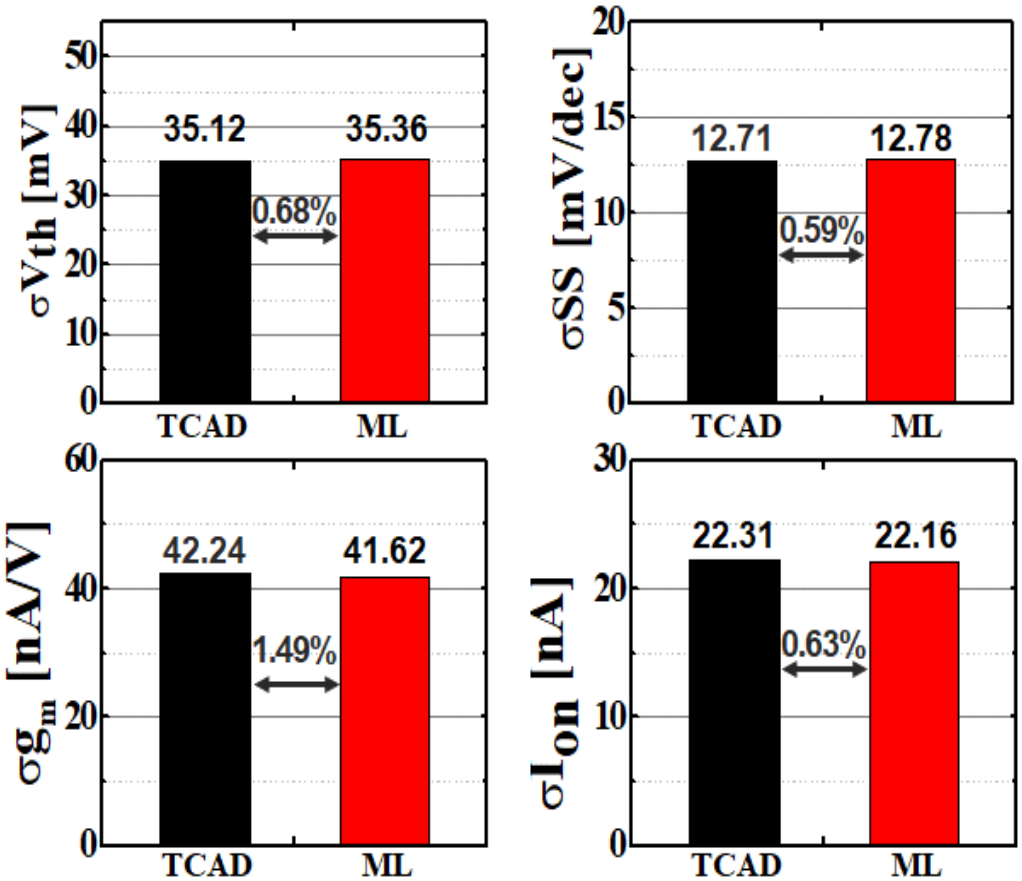


Fig. 4.8. The results of TCAD simulation and ML algorithm for standard deviation of key electrical parameters. The average error rate is less than 1%.

4.3.2 Taper Angle Effect with Various Stacked Layers of 3D NAND Flash Memories

The taper angle, a problem inevitably affecting the vertical device that occurs during the channel etching process, can affect the electrical characteristics of the cells in the string [9, 10]. Therefore, unlike the other seven variability sources mentioned above, the effects of the tapered channel can be observed in strings rather than in a single word line.

Figure 4.9 shows the channel cross-sections and Id-Vg characteristics due to tapered channel effects. The change in channel cross-sectional area along the taper angle is a problem that cannot be controlled in the process, so the lower cell has a relatively narrower channel area than the upper cell of the string. Therefore, the average channel cross-sectional area of all of the cells existing in the string is reduced as the number of layers of the memory devices and the taper angle increase. In other words, as the number of stacked layers of the 3D NAND Flash memory device increases, the cell located at the bottom of the string may disappear even with a smaller tapered angle. In addition, the decreasing channel area as the cell is positioned below reduces the operating current of the entire channel. As a result, it can be a major factor that degrades the overall electrical characteristics. Therefore, the effects of various variability sources can be further increased due to the reduced channel region. In addition, it is assumed that the change in channel cross section is equally reduced without considering the aspect ratio.

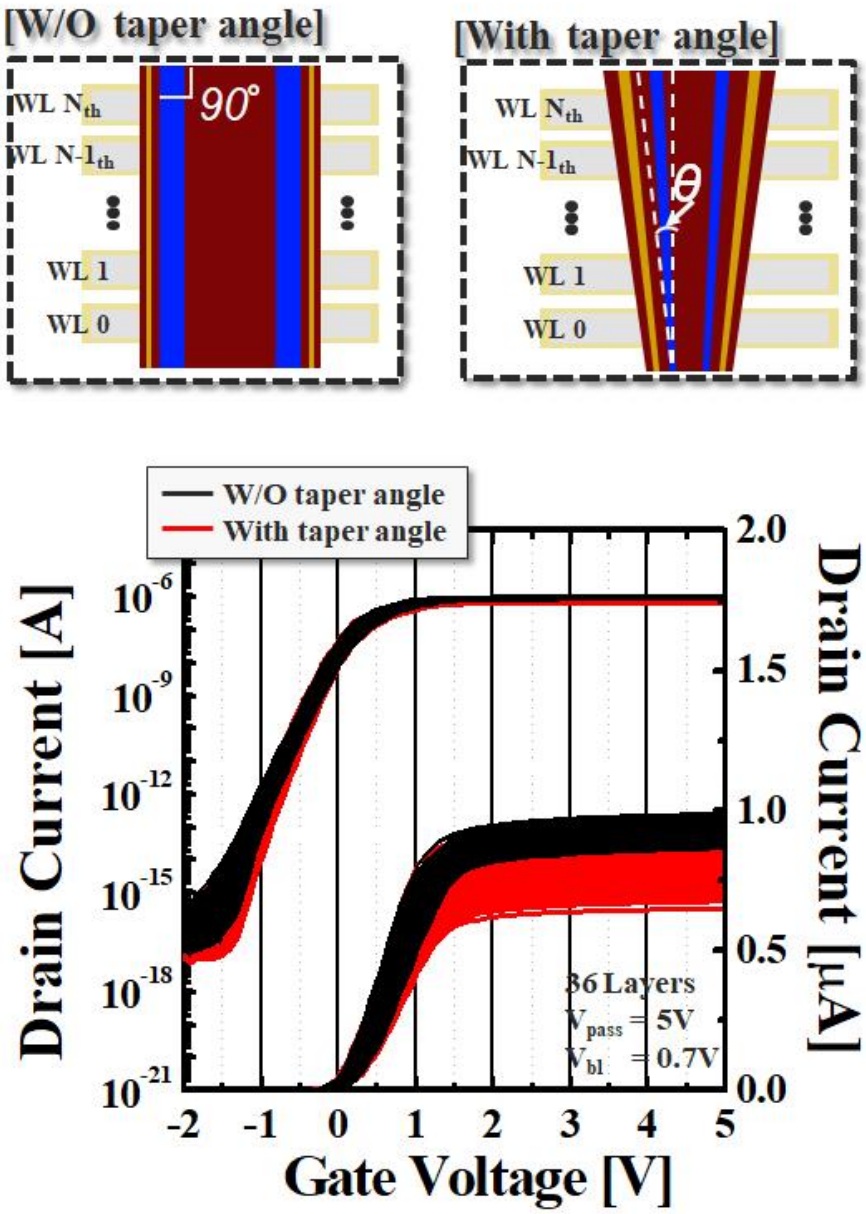


Fig. 4.9. A cross section of a 3D NAND Flash memory device containing a taper angle. Id-Vg plot of 600 3D NAND Flash memory devices with and without taper angle.

Figure 4.10 shows the variation in the electrical characteristic distribution with increasing taper angle in the 36-layer memory devices. As the taper angle increases, the average value (μ) of the key electrical characteristic distribution deteriorates due to the decreased average channel cross-sectional area. In addition, it can be seen that the effects of variability sources are increased even when considering the factor of the same parameter.

Figure 4.11 shows the distributions and the correlation between changes in electrical characteristics caused by variability sources due to tapered channel effects. Because of the simultaneous effects of various variability sources, the correlation between the electrical characteristic distributions is low due to the effect of independent variability sources. Also, many cells exist in a single string, and the effects of GBT and CD control affect the entire string. Therefore, the distribution of electrical characteristics of the selected cell is a result including the variation effect of other cells existing in the same string. In other words, since the conditions of both the selected cell and the neighboring cells are considered, the electrical characteristic distributions extracted from the selected cell may show a low correlation. Nevertheless, the results show that the distribution according to this characteristic change also predicts the ANN-based ML algorithm very accurately. In addition, the σ/μ bar graph of the two distributions shows that the rate of error is close to nil for all key electrical parameter variations.

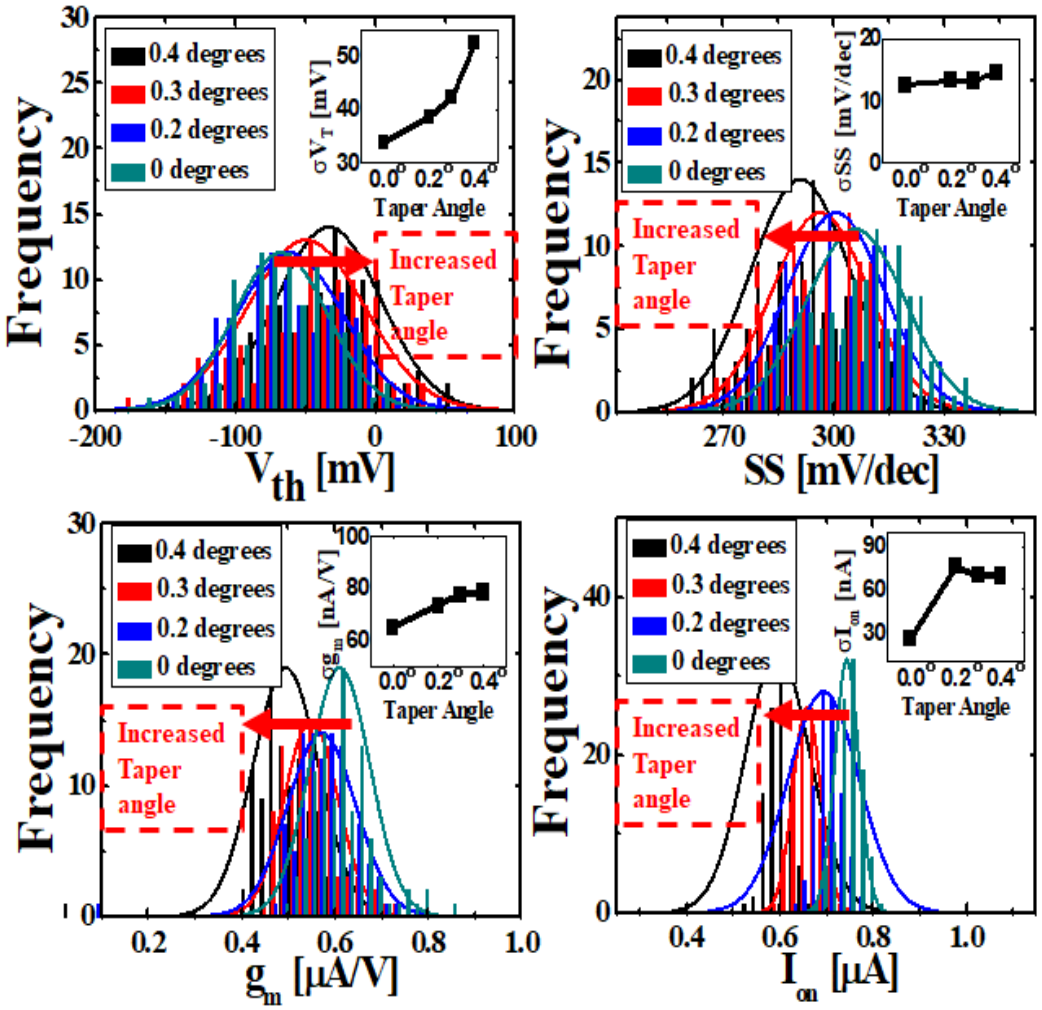


Fig. 4.10. The shift of the key electrical parameters distribution of 3D NAND Flash memory devices with increase of taper angle. The illustrations in each figure show an increase in the standard deviation of the key electrical parameter distribution with increasing taper angle.

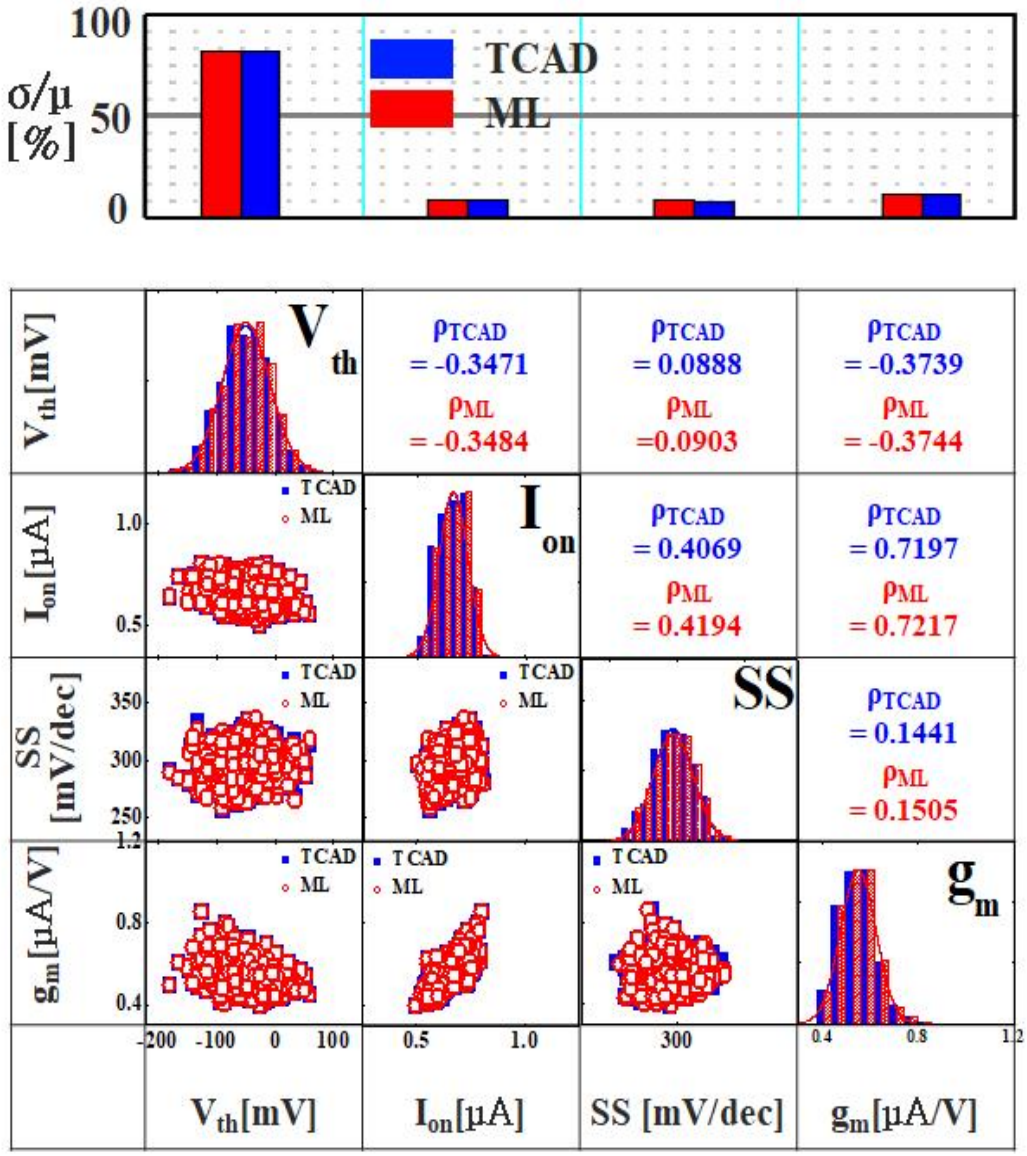


Fig. 4.11. Comparison of scatter plots and correlations of the key electrical parameters between the TCAD data and the ML Approach results from 3D NAND Flash memory devices.

Figure 4.12 shows the effect of variability sources on the increase in the number of layers of 3D NAND flash memory. With an increasing number of layers, the tapered channel effect increases the influence of variability sources. In addition, as the number of layers increases, the influence increases even in the case of having the same channel angle, so that the average value of the electrical characteristic distribution is changed. Increasing the influence of the tapered channel effect is closely related to the decrease of the relative channel cross-sectional area. Therefore, due to the reduction of the channel area, the average value of the threshold voltage is increased, the operating current and the transconductance are decreased, and the SS can be slightly improved. The maximum taper angle was set based on 64-layer memory devices for quantitative comparison. When the stacked stage of the memory device is fixed at 64 stages, the maximum channel tapered angle that can be formed is about 0.3 degrees. If the stack has more than 64 stages, the channel tapered angle should be set smaller than 0.3 degrees to form the channel region in the cell located at the bottom of the string. Therefore, when considering the structural change of the 3D NAND flash device due to the increase of the number of layers, it can be confirmed that the ANN-based ML algorithm accurately predicts the effects of variability sources, including the tapered channel effects. In order to evaluate the effect of the variability source, the data for training (about 100 sample devices) and test (about 600 sample devices) are independently extracted and used through TCAD simulation.

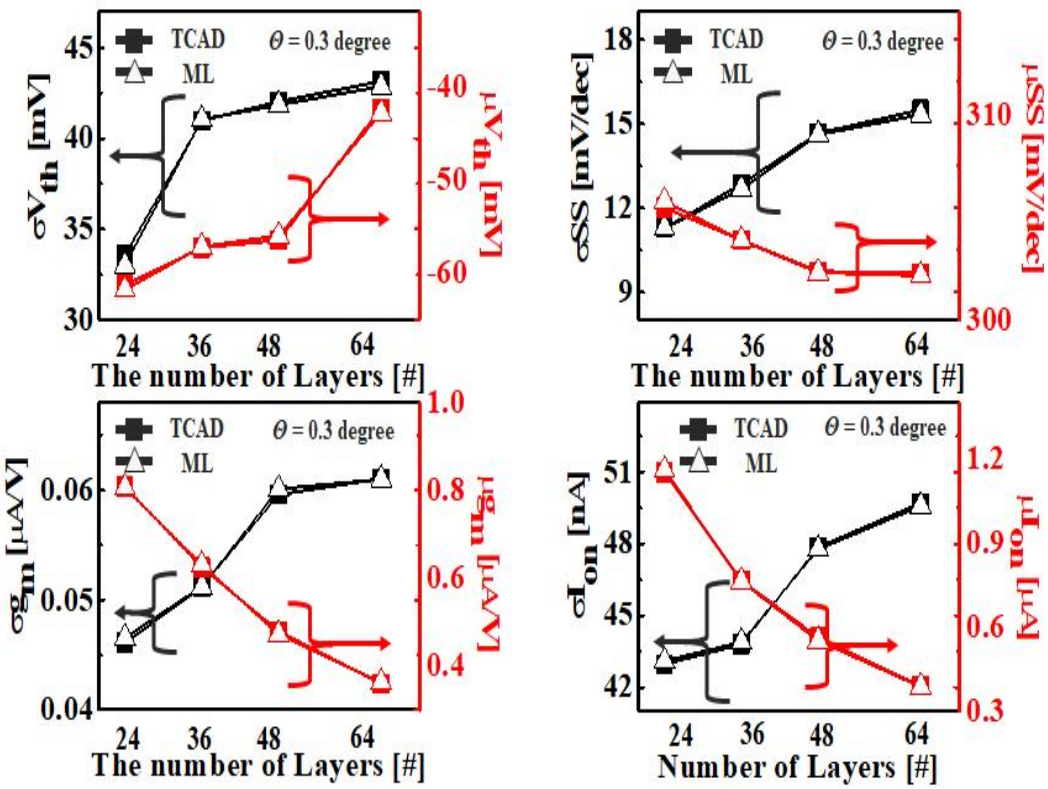


Fig. 4.12. The variation of the mean and standard deviation of the distribution of the key electrical parameters as the number of stacked layers of 3D NAND Flash memory devices increases.

4.3.3 Statistical Analysis of 3D NAND Flash Memory Device Program and Erase Operation

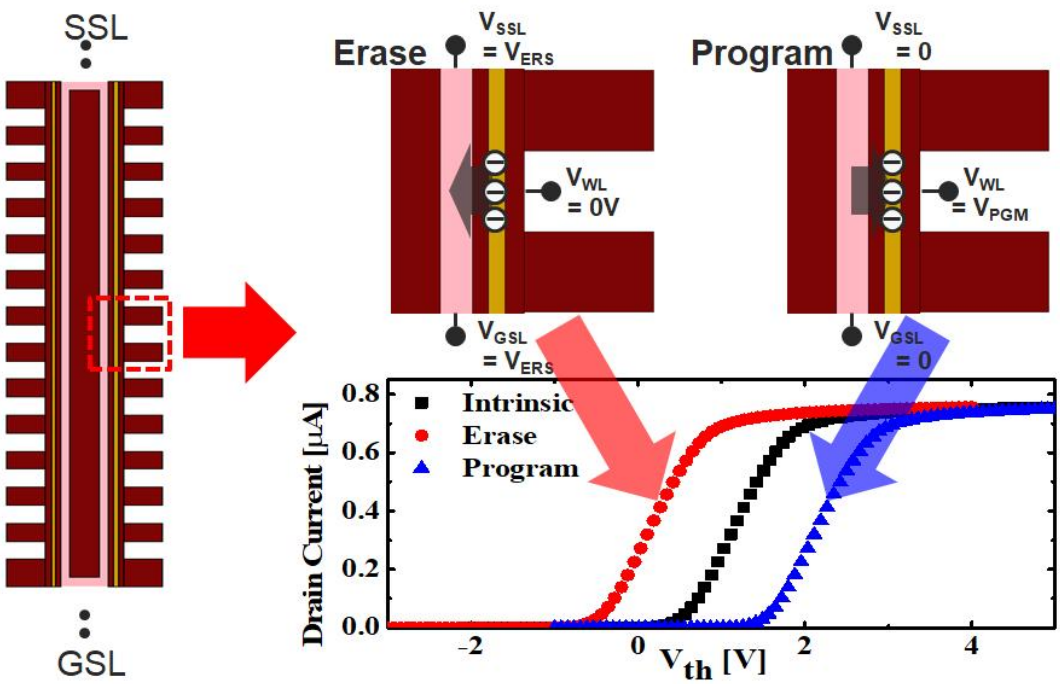


Fig. 4.13. Schematic of program / erase operation of 3D NAND Flash memory device.

The basic operation principle of 3D NAND flash memory is Erase/Program/Read. Since NAND Flash is a non-volatile memory, the desired data must be correctly read for a certain period of time even if the stored data is turned off. The data program and erase

method stores / erases electrons by tunneling the oxide layer in the electron storage layer (charged layer, Nitride) by applying a high bias. The number of electrons stored in the charged layer can be adjusted through various voltage conditions, and the cell Threshold voltage is determined according to the number of stored electrons. During the program operation, high voltage is applied to the control gate to store the electrons in the charged layer. As more electrons are stored in the charged layer, the Threshold voltage of the device increases. NAND Flash is classified into SLC (Single Level Cell) with 1 bit information, MLC (Multi Level Cell) with 2 bits information, TLC (Triple Level Cell) with 3 bits information, and QLC (Quad Level Cell) with 4 bits information. As more information is stored in a cell, the number of Threshold voltage classifications of cells increases as SLC → MLC → TLC → QLC goes in order to distinguish each information by differently storing the number of electrons. In the case of erase operation, since a high voltage is applied to the substrate region, electrons stored in the charged layer are removed to reduce the Threshold voltage of the device. During the read operation, a relatively low voltage is applied to the control gate rather than during programming to distinguish where the cell Threshold voltage is located based on the read level, whether it is a program cell or an erase cell through string-current sensing. The retention characteristic is a very important reliability item indicating how long the stored electrons can be read accurately because NAND Flash is a non-volatile memory.

Figure 4.13 explains the program and erase operation principle of a 3D NAND Flash memory device. The number of trapped charge in the charged layer changes according to

the applied voltage condition. Therefore, by analyzing the Id-Vg characteristics of the device, it can be confirmed that Threshold voltage changes as the program and erase operations are performed compared to the intrinsic state.

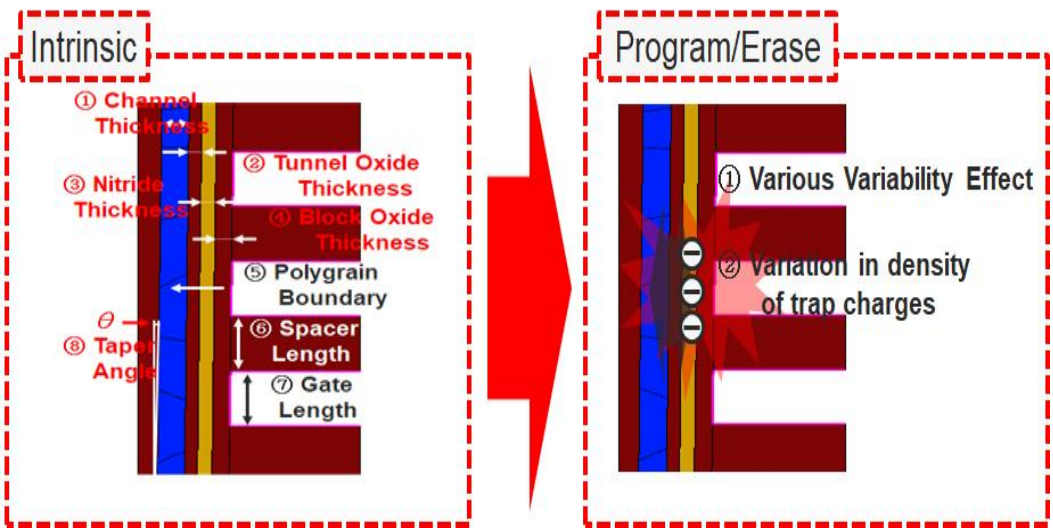


Fig. 4.14. Variation sources in the intrinsic state and changes in the number of trap charges due to program / erase operation

Figure 4.14 shows the variability sources analyzed in the device in the intrinsic state and the change in the number of trap charges according to the program and erase operations. The variability sources analyzed in the intrinsic state are factors that independently change the characteristics of the device. However, in the case of a change in the number of trap charges, it is dependently affected by the aforementioned variability source. For example, among the variability sources in the intrinsic state, the change in

channel thickness and the change in oxide film thickness are independent of each other. The fact that the thickness of the channel is formed thinly does not affect the thickness of the oxide film. However, the number of trap charges due to program and erase operations is a variable factor that is directly affected by the thickness of the oxide film or the thickness of the channel. Therefore, the Threshold voltage distribution according to the intrinsic variability sources may be analyzed in advance, and then the variation of the device characteristics according to the program and erase operation may be analyzed and predicted by setting the variation of the trap charge.

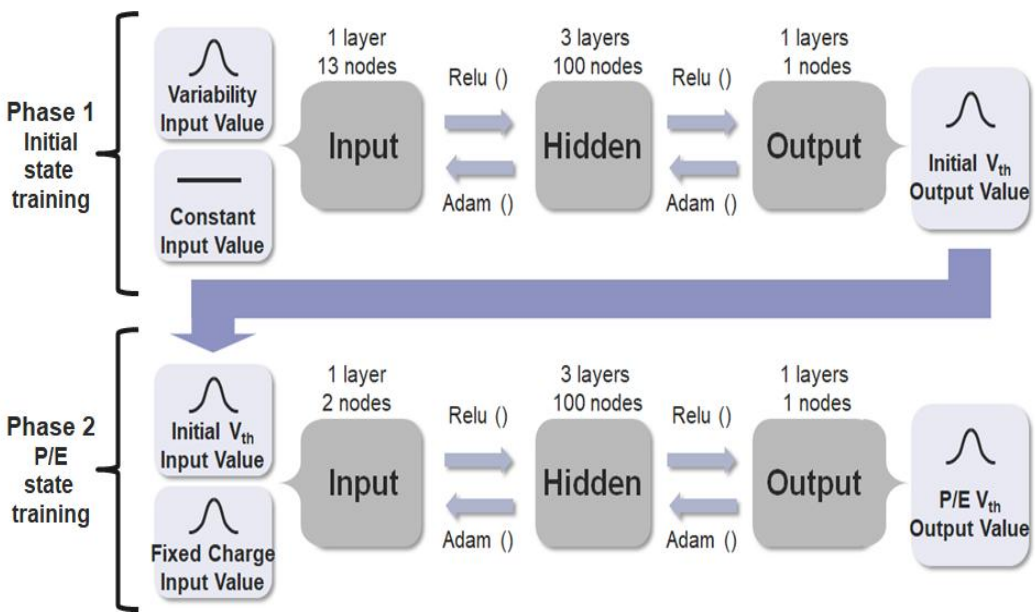


Fig. 4.15. Schematic diagram of the algorithm for predicting the threshold voltage distribution according to the variability sources of the intrinsic state and for predicting the threshold voltage distribution according to the P / E operation.

Figure 4.15 shows the steps of predicting the variation of the threshold voltage of the device according to the P / E operation based on the structure of the machine learning system based on the ANN algorithm used to predict the variability factor of the initial state. In the first phase algorithm, the Threshold voltage distribution according to the variability source of the intrinsic state is predicted. Therefore, the machine learning system is constructed based on the preceding results without changing the previous algorithm. In the second phase algorithm, the Threshold voltage distribution in the initial state is received as a new input value. In addition, a change in the number of trap charges according to the program and erase operation is received as an input value. As a result, an algorithm for receiving the trap charge amount dependent on the intrinsic variability source as a new input value is continuously set. Therefore, the final Threshold voltage distribution is predicted to include the effect of the variability source in the intrinsic state and the effect of the change in the trap charge according to the program/erase operation. By using this dual algorithm, this study can effectively predict the threshold voltage distribution of a 3D NAND flash memory device.

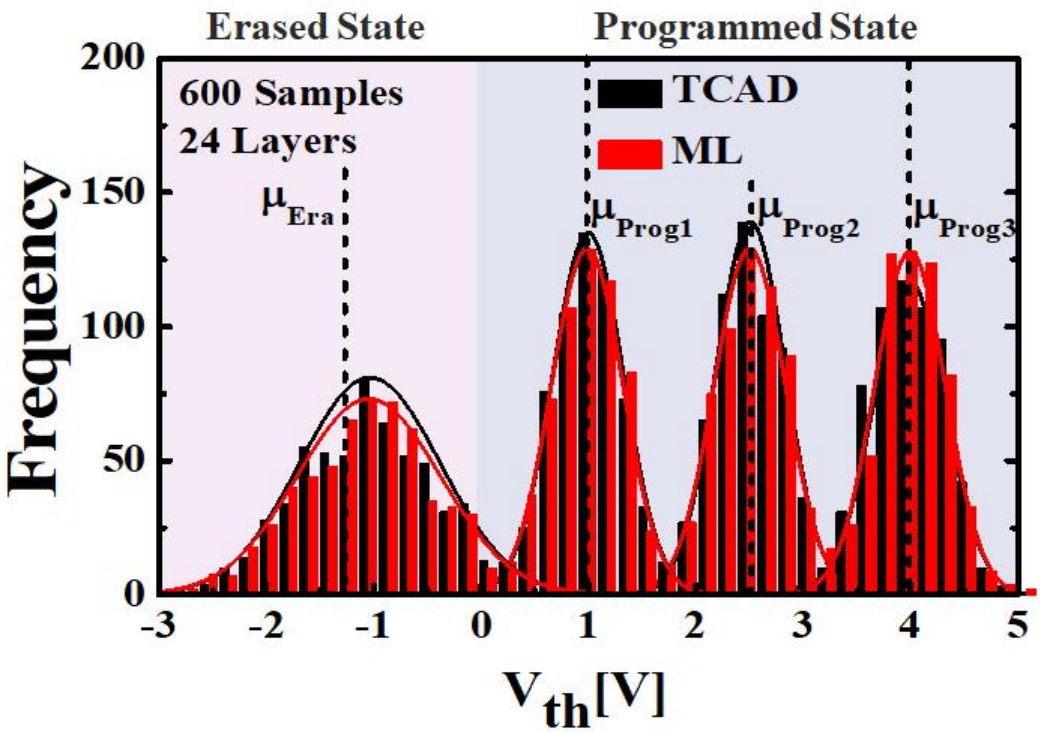


Fig. 4.16. Statistical comparison of TCAD simulation and machine learning results for 3D NAND Flash memory device program and erase operations.

Unlike the initial state, when various variability sources are affected, the threshold voltage distribution is shown in Figure 4.16 during program and erase operations, which are the basic operations of the memory device. The standard deviation error rate of the threshold voltage distribution using the TCAD simulation and ML approach differs by an average of about 2%, and it can be seen that the prediction is very good even when using independent test data, as shown in table 4.4. For comparison of program and erase

operations, the TCAD simulation values were corrected for the measured values, referring to the previous study [20, 21]. Compared to the initial state, the distribution of threshold voltages increases during program and erase operations. To implement this, simulation is set to randomly distribute the fixed charge in the charge trap layer (Nitride Layer). Therefore, in order to train and test the extracted threshold voltage distribution through the ML system, a fixed charge value is added as a new input variable to the existing algorithm. Therefore, the prediction result can be extracted without changing the existing algorithm.

Performance Parameters								
V_{th} Distribution [V]								
	Erase		Program 1		Program 2		Program 3	
	TCAD	ML	TCAD	ML	TCAD	ML	TCAD	ML
Sigma[σ]	0.6539	0.6674	0.3301	0.334	0.3221	0.3331	0.3366	0.3309
Average[μ]	-1.0358	-1.0467	0.9956	0.9871	2.5083	2.5006	3.9874	3.9926
Err [RMS (%)]	0.0135(2.02)		0.004(1.19)		0.0109(3.29)		0.0057(1.71)	

Table 4.4. Summary of threshold voltage distribution prediction according to Program / erase operation of 3D NAND flash device.

4.3.4 Efficiency and Accuracy of Machine Learning Systems.

The figure below shows a quantitative approach for training and testing ANN algorithms in stages. In addition, in this study, we use a configuration of the computer: Intel Xeon CPU E5-2687W 216 V2 (3.5 GHz × 16) processor with 128 GB of RAM, which is same with our previous work [17]. All data used in the study were benchmarked using calibrated TCAD simulation data based on measurement data. Therefore, it is possible to calculate the quantitative calculation cost by operating the TCAD simulation and the machine learning system in the same server environment.

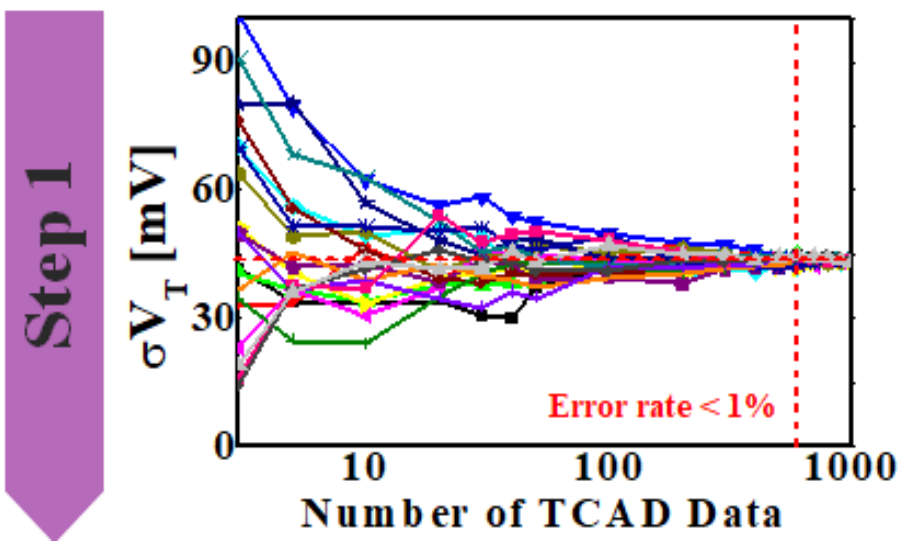


Fig. 4.17. Extract the minimum amount of data for the distribution of electrical characteristics of a 3D NAND Flash memory device to a normal distribution

In the first step, the same as the approach in section 3, we determine the number of random device samples that are statistically required when analyzing TCAD simulations independently. The variability source used in this study are analyzed inclusively of local and global variability sources. Thus, the data is corrected in a form suitable for insertion into a ML system through data preprocessing. Even if data preprocessing is performed, all the variability sources are converged to Gaussian distribution if enough data is extracted. Thus, in order to understand the characteristic change of the device caused by such a variable source, a minimum number of sample data is required to form a characteristic change such as a Gaussian distribution. As a result, similar to the results in section 3, when analyzing the characteristic change in more than about 600 random devices, it can be seen that the standard deviation converges to a specific Gaussian distribution. In other words, to statistically analyze the variability source, at least 600 TCAD simulation device analyzes must be preceded. In the process, TCAD simulations can be used to calculate the cost of calculations for analyzing the effects of variability.

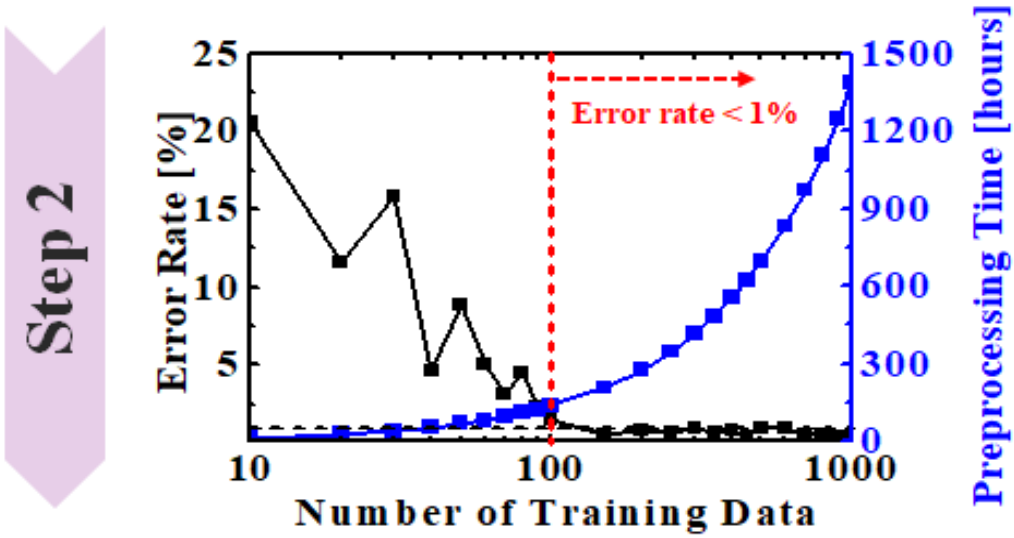


Fig. 4.18. Extract the minimum amount of training data to achieve the same prediction accuracy as the TCAD simulation.

In the second step, the data extracted from the TCAD simulation data is used to quantitatively analyze the minimum amount of training data to train the ML system. For ML systems using the ANN algorithm, a large amount of data samples are not needed to form a Gaussian distribution in order to fully train the compute nodes inside the algorithm. Since the training of the algorithm proceeds to calculate the correct output value for the randomly extracted data, a small amount of data for training is possible. Therefore, the minimum number of random device samples for learning the ANN algorithm presented in this study is about 100. As the amount of learning data increases, the predictive accuracy of ML systems increases.

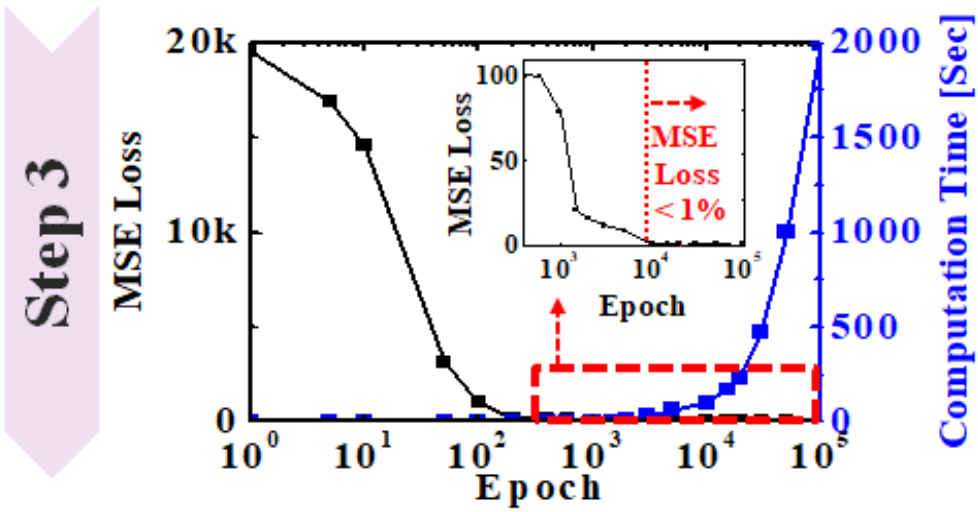


Fig. 4.19. Number of iterations and time required for internal algorithms to train machine learning systems.

In the third stage, the minimum number of repetitive training sessions for sufficient training is analyzed through about 100 training data sets defined in the previous stage. The ANN algorithm adjusts the weighting value assigned to each node through iterative calculation of the activation function and the optimization function, as described in section 2, for sufficient formation of internal computational nodes. Therefore, when the training proceeds beyond the minimum number of repetitive training, the error rate of the internal algorithm such as MSE Loss gradually decreases. In other words, this study shows that the error rate is effectively reduced when the number of repetitions of the ML system to accurately predict the set variability source is performed about 20,000 times.

Table 4.4 quantitatively shows this prediction accuracy. The mean (μ) and standard

deviation (σ) of the key electrical characteristic distributions of various 3D NAND flash memories caused by various variability sources are extracted by 3D TCAD simulation and ANN-based ML algorithm. In all cases shown in Table 4.5, the test data for comparison are extracted and compared independently from the training data. The results show that the error between the two methods is less than 1% on average.

In terms of computational cost, the following steps are used to quantitatively compare the two approaches. First, the process of extracting a sufficient number of sample device data through a 3D TCAD simulation to form a normal distribution (requires about 600 or more sample data). Extract the minimum required training data for proper training of the ML system (approximately 15% data sample required). Finally, the training and testing phase of the trained ML system is performed (required calculation cost less than 0.01% compared to 3D TCAD simulation). As a result, the total computational cost of the ML system is only about 16% compared to 3D TCAD simulation. In addition, with error rate verification through independent test data, the accuracy of the ML system is equivalent to benchmarking TCAD simulation results. Thus, ML systems are more efficient than conventional approaches and can achieve the same level of accuracy.

Performance Parameters												
No. of Layer	V_{th} [mV]						SS [mV/dec]					
	σ_{TCAD}	σ_{ML}	Err [RMS (%)]	μ_{TCAD}	μ_{ML}	Err [RMS (%)]	σ_{TCAD}	σ_{ML}	Err [RMS (%)]	μ_{TCAD}	μ_{ML}	Err [RMS (%)]
24	33.57	33.01	0.56 (1.67)	-61.25	-61.77	0.52 (0.85)	11.28	11.2709	0.01 (0.09)	305.46	305.78	0.3187 (0.1)
36	40.98	41.05	0.07 (0.17)	-57.29	-57.18	0.11 (0.19)	12.85	12.6759	0.1803 (1.4)	303.89	303.87	0.0158 (0.01)
48	42.03	41.85	0.18 (0.43)	-56.46	-55.99	0.47 (0.83)	14.66	14.6159	0.0461 (0.31)	302.37	302.31	0.0589 (0.02)
64	43.16	42.87	0.29 (0.67)	-42.37	-42.87	0.5 (1.18)	15.49	15.3258	0.1725 (1.11)	302.28	302.15	0.1231 (0.04)
No. of Layer	g_m [μA/V]						I_{on} [μA]					
	σ_{TCAD}	σ_{ML}	Err [RMS (%)]	μ_{TCAD}	μ_{ML}	Err [RMS(%)]	σ_{TCAD}	σ_{ML}	Err [RMS (%)]	μ_{TCAD}	μ_{ML}	Err [RMS (%)]
24	0.0461	0.0466	0.0005 (1.08)	0.8079	0.8055	0.0023 (0.29)	0.043	0.0431	0.0001 (0.33)	1.1511	1.1586	0.0076 (0.66)
36	0.0512	0.0512	0 (0.04)	0.6243	0.63	0.0057 (0.92)	0.0438	0.0439	0.0001 (0.18)	0.7669	0.7626	0.0043 (0.56)
48	0.0596	0.0601	0.0005 (0.87)	0.4789	0.4713	0.0076 (1.6)	0.0479	0.0478	0.0001 (0.19)	0.5629	0.5519	0.0111 (1.96)
64	0.0611	0.061	0.0001 (0.13)	0.3592	0.3625	0.0032 (0.9)	0.0497	0.0496	0.0001 (0.26)	0.3905	0.3876	0.003 (0.76)

Table 4.5. Summary of the optimization results on a 3D NAND Flash memory devices.

4.4 Summary

For the first time, we proposed an ANN-based ML approach to predict and analyze the effects of various variability sources in 3D NAND flash memories. The 3D stochastic TCAD simulation, which is commonly used to simultaneously analyze various variability sources of 3D NAND flash memory, has high accuracy in terms of prediction and analysis but also has a very high computational cost. In addition, when considering the structure of 3D NAND Flash memory devices, string and block-level analysis are more important than a single cell. In other words, simultaneous analysis through large sample data is essential, which can further increase the cost of computation. In order to overcome this problem, the ANN-based ML system can be a new method to accurately predicting various variability sources in a large sample of 3D NAND Flash memory device data. The proposed ML system requires only about 15% of the minimum sample data for proper algorithm training. Nevertheless, the proposed ML system achieves a level of accuracy equivalent to that of the conventional approach and promotes significant efficiency improvements. As a result, our ANN-based ML approach can significantly improve (6×) the calculation efficiency while maintaining the high accuracy (relative error of approximately 1%) of the existing 3D stochastic TCAD simulation.

References

[1] D. Resnati, A. Mannara, G. Nicosia, G. M. Paolucci, P. Tessariol. A. S. Spinelli, A. L. Lacaita, and C. M. Compagnoni, " Characterization and Modeling of Temperature," IEEE Trans. Electron Devices, Vol. 65, No. 8, pp. 3199-3206, Aug. 2018, doi: 10.1109/TED.2018.2838524

[2] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y,. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, "Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory" in Proc. Symp. VLSI Technol., pp. T14-T15, Jun. 2007, doi: 10.1109/VLSIT.2007.4339708.

[3] H. Aochi, "BiCS Flash as a Future 3D Non-Volatile Memory Technology for Ultra High Density Storage Devices" in proc. IEEE Int. Memory Workshop, pp. 1-2, May. 2009, doi: 10.1109/IMW.2009.5090581

[4] M. K. Jeong, S. M. Joe, C. S. Seo, K. R. Han, E. Choi, S. K. Park, and J. H. Lee, "Analysis of Random Telegraph Noise and low frequency noise," in Proc. Symp. VLSI Technol., pp. 55-56, Jul. 2012, doi: 10.1109/VLSIT.2012.6242458.

[5] A. S. Spinelli, C. M. Compagnoni, and A. L. Lacaita, "Reliability of NAND Flash Memories: Planar Cells and Emerging Issues in 3D Devices", Computers, Vol. 6, No. 2, pp. 1-55, Mar. 2017, doi: 10.3390/computers6020016.

[6] Y. Yanagihara, K. Miyaji, and K. Takeuchi, "Control Gate Length, Spacing and

Stacked Layer Number Design for 3D-Stackable NAND Flash Memory" in proc. IEEE Int. Memory Workshop, pp. 1-4, May. 2012, doi: 10.1109/IMW.2012.6213656

[7] B. Kim, S. H. Lim, D. W. Kim, T. Nakanishi, S. Yang, J. Y. Ahn, H. M. choi, K. Hwang, Y. Ko, and C. J. Kang, "Investigation of Ultra Thin Polycrystalline Silicon Channel for Vertical NAND Flash" IEEE Int. Reliability Physics Symposium, pp. 2E.4.1-2E.4.4, Apr. 2011, doi: 10.1109/IRPS.2011.5784464.

[8] R. Degraeve, S. Clima, V. Putcha, B. Kaczer, Ph. Roussel, D. Linten, G. Groeseneken, A. Arreghini, M. Karner, C. Kernstock, Z. Stanojevic, G. V. D. Bosch, J. V. Houdt, A. Furnemont, and A. Thean, "Statistical Poly-Si grain boundary model with discrete charging defects and its 2D, IEEE Int. Electron Devices Meeting, Dec. 2015, pp. 5.6.1-5.6.4, doi: 10.1109/IEDM.2015.7409636.

[9] Y. Oh, K. B. Kim, S. H. Shin, H. Sim, N. V. Toan, T. Ono, and Y. H. Song, "Impact of etch angles on cell characteristics in 3D NAND flash memory" Microelectronics Journal, Vol. 79, pp. 1-6, Sep. 2018, doi: 10.1016/j.mejo.2018.06.009

[10] K. T. Kim, S. W. An, H. S. Jung, K. H. Yoo, and T. W. Kim, "The Effects of Taper-Angle on the Electrical Characteristics of Vertical NAND Flash Memories" IEEE Electron Device Letters, Vol. 38, No. 10, pp. 1375-1379, Oct. 2017, doi: 10.1109/LED.2017.2747631.

[11] T. Ohashi, A. Yamaguchi, K. Hasumi, M. Ikota, G. Lorusso, C. L. Tan, G. V. D. Bosch, and A. Furnemont, "Precise measurement of thin-film thickness in 3D-NAND device with CD-SEM" Journal of Micro/Nanolithography MEMS and MOEMS, Vol. 17,

No. 2, May. 2019, doi: 10.1117/1.JMM.17.2.024002.

[12] M. Abe, T. Nakamura, and K. Takeuchi, “Pre-shipment Data-retention/ Read-disturb Lifetime Prediction & Aftermarket Cell Error,” in Proc. Symp. VLSI Technol., pp. T216-T217, Jun. 2019, doi: 10.23919/VLS IT.2019.8776480.

[13] G. W. Burr, R. M. Shelby, S. Sidler, C. D. Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, “Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses),” IEEE Trans. Electron Devices, Vol. 62, No.11, pp. 3498-3507, Nov. 2015, doi: 10.1109/TED.2015.2439635

[14] J. Virarahgavan, S. J. Pandharpure, and J. Watts, “Statistical Compact Model Extraction:,” IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, Vol. 31, No. 12, pp. 1920-1924, Dec. 2012, doi: 10.1109/TCAD.2012.2207955.

[15] V. Janakiraman, A. Bharadwaj, and V. Visvanathan, “Voltage and Temperature Aware Statistical Leakage,” IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, Vol. 29, No. 7, pp. 1056-1069, Jul. 2010, doi: 10.1109/TCAD.2010.2049059.

[16] V. Subramanian, Deep Learning with Pytorch, 1st ed. iG Publishing Pte. Ltd. Press, 2018.

[17] K. Ko, J. K. Lee, M. Kang, J. Jeon, and H. Shin, “Prediction of Process Variation Effect for Ultrascaled GAA Vertical FET Devices Using a Machine Learning Approach” IEEE Trans. Electron Devices, Vol. 66, No.10, pp. 4474-4477 Sep. 2019, doi: 10.1109/TED.2019.2937786

[18] ITRS, Denver, CO, USA. International Technology Roadmap for Semiconductors(ITRS), 2015. Online(http://www.itrs2.net/)

[19] Sentaurus Device User Guide Version: K_2015.06, Synopsys, Mountatin View, CA, USA, Jun. 2015.

[20] D. Lee, and W. Sung, “Estimation of NAND Flash Memory Threshold Voltage Distribution for Optimum Soft-Decision Error Correction”, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 61, NO. 2, Jan. 2013, doi: 10.1109/TSP.2012.2222399

[21] H. Li, “Modeling of Threshold Voltage Distribution in NAND Flash Memory: A Monte Carlo Method”, IEEE Trans. Electron Devices, VOL. 63, NO. 9, Sep. 2016, doi: 10.1109/TED.2016.2593913

5. Conclusion

In this dissertation, we propose a new approach to analyze and predict the effects of variability sources on various devices through ANN algorithm-based ML systems. This new ML system has significantly improved efficiency and scalability compared to traditional analytical and predictive methods.

Chapter 1 describes the issues of variability caused by the development of next generation semiconductor devices. Unlike in the past, the difficulty of interpreting variability sources increases because it reauires integrated analysis of various variability sources. Therefore, we introduce an ML system to meet the needs of new systems that go beyond traditional approaches.

Chapter 2 describes in detail how to integrate with TCAD simulation and algorithm implementation for ML system. This section provides an in-depth introduction to the various issues arising from the use of ML systems, and how to configure an optimal system. In addition, this section introduces the development of ML systems with connectivity on various platforms, aimed at interoperability with commonly used TCAD simulations.

Chapter 3 analyzes and predicts the effects of variability sources on ultra-scaled GAA devices through ML systems. The method of initial analysis through the ML system accurately predicts the distribution of changes in electrical characteristics due to various

varability sources. Moreover, it can be seen that the efficiency is improved by presenting the quantitative comparison method as compared with the conventional method. The method through the ML system is very efficient for the integrated analysis of the effects of various variability sources.

Chapter 4 introduces the advanced ML system of the previous logic device. This section shows an ML system that simultaneously analyzes and predicts various variability sources of 3D NAND Flash memory. The proposed ML system predicts the change in electrical characteristic caused by covering all stacking layer and process variation source of 3D NAND Flash memory device. Moreover, quantitative comparison with TCAD simulation confirms the accuracy, generality and efficiency of the ML system.

As a result, the method of predicting the effects of various variability sources proposed in this paper can provide useful guidelines to those who develop and design 3D NAND Flash memory devices.

Bibliography

Journal

[1] Kyul Ko, Jangkyu Lee, and Hyungcheol Shin, “Variability-Aware Machine Learning Strategy for 3D\biotitnand Flash Memories” IEEE TED, 2020.

[2] **Kyul Ko,** Myounggon Kang, Jeon Jongwook, and Hyungcheol Shin, “Compact model strategy of metal-gate work-function variation for ultrascaled FinFET and vertical GAA FETs” IEEE TED, 2019.

[3] **Kyul Ko,** Myounggon Kang, Jeon Jongwook, and Hyungcheol Shin, “Variability-Aware Simulation Strategy for Gate-All-Around Vertical Field Effect Transistor”, Journal of Nanoscience and Nanotechnology, 2019.

[4] **Kyul Ko,** Jangkyu Lee, Myounggon Kang, Jeon Jongwook, and Hyungcheol Shin, “Prediction of Process Variation Effect for Ultrascaled GAA Vertical FET Devices Using a Machine Learning Approach.”, IEEE TED, 2019.

[5] **Kyul Ko,** Dokyun Son, Myounggon Kang, and Hyungcheol Shin, “Comparison of work function variation between FinFET and 3D stacked nanowire FET devices for 6-T SRAM reliability.”, Solid-State Electronics, 2018.

[6] **Kyul Ko**, Dokyun Son, Myounggon Kang and hyungcheol Shin, “Analysis and Comparison of Interface Trap For Single and 3D Stacked Nanowire FET” Journal of Nanoscience and Nanotechnology, 2017.

[7] **Kyul Ko**, Changbeom Woo, Minsoo Kim, Youngsoo Seo, Shinkeun Kim, Myounggon Kang, and Hyungcheol Shin, “Analysis and Comparison of Intrinsic Characteristics for Single and Multi-channel Nanoplate Vertical FET Devices”, JSTS, 2017.

[8] **Kyul Ko**, Dokyun Son, Changbeom Woo, Myounggon Kang and hyungcheol Shin, “Investigation of Work Function Variation Induced by Metal Gate and Process Variation Effect in 3D Stacked Nanowire FET Devices” Journal of Nanoscience and Nanotechnology, 2017

[9] Youngsoo Seo, Shinkeun Kim, **Kyul Ko**, Changbeom Woo, Minsoo Kim, Jangkyu Lee, Myounggon Kang, and Hyungcheol Shin, “Analysis of electrical characteristics and proposal of design guide for ultra-scaled nanoplate vertical FET and 6T-SRAM:” Solid-State Electronics, 2018

[10] Dokyun Son, **Kyul Ko**, Myounggon Kang and hyungcheol Shin, “3D Technology Computer-Aided Design-Based Optimization of Channel Radius Considering Line Edge Roughness on Gate-All-Around Nanowire FET” Journal of Nanoscience and Nanotechnology, 2017

[11] Dokyun Son, **Kyul Ko**, Changbeom Woo, Myounggon Kang, and hyungcheol Shin, “Characteristics According to Parameters of Line Edge Roughness in Ultra-Scaled Gate-All-Around Nanowire FET” Journal of Nanoscience and Nanotechnology, 2017.

[12] Dokyun Son, **Kyul Ko**, Changbeom Woo, Myounggon Kang, and hyungcheol Shin, “Line Edge Roughness and Process Variation Effect of Three Stacked Gate-All-Around Silicon MOSFET Devices” Journal of Nanoscience and Nanotechnology, 2017

[13] Changbeom Woo, **Kyul Ko**,, Jongsu Kim, Minsoo Kim, Myounggon Kang and Hyungcheol Shin, “Analysis and Optimization of RC Delay in Vertical Nanoplate FET,” Solid-State Electronics, 2017.

Conference

[1] **Kyul Ko**, Yeaji Yoo, Myunggon Kang, Jongwook Jeon, and Hyungcheol Shin, “Variability-Aware Simulation of Tapered GAA Vertical FETs” Korean Conference on Semiconductors, 2019.

[2] **Kyul Ko**, Myunggon Kang, and Hyungcheol Shin, “Modeling and Analysis of Work Function Variation in Nanowire FET”, Korean Conference on Semiconductors, 2018

[3] **Kyul Ko**, Dokyun Son, and Hyungcheol Shin, “Analysis of Metal gate Work-Function Variation for Vertical Nanoplate FET in 6-T SRAMs”, Silicon Nanoelectronics Workshop, 2017.

[4] **Kyul Ko,** Changbeom Woo, Minsoo Kim, Youngsoo Seo, Shinkeun Kim, Myounggon Kang, and Hyungcheol Shin, “Analysis and Comparison of Intrinsic Characteristics for Single and Multi-channel Nanoplate Vertical FET Devices” Int'l Conference on Semiconductor Physics and Devices, 2017.

[5] **Kyul Ko,** Dokyun Son, Changbeom Woo, Myounggon Kang and hyungcheol Shin, “Analysis of 6T SRAM performance by Work Function Variation for 3D Stacked Nanowire FET Devices” Korean Conference on Semiconductors, 2017.

[6] **Kyul Ko,** Dokyun Son, Myounggon Kang and hyungcheol Shin, “Analysis of Work Function Variation and Process Variation effect for 5nm node 3D Stacked Nanowire FET Devices” International Vacuum Congress, 2016

[7] **Kyul Ko,** Dokyun Son, Myounggon Kang and hyungcheol Shin, “ Analysis and Comparison of Work Function Variation for 5nm node Single Nanowire FET and 3D Stacked Nanowire FET Devices” NanoKorea, 2016

[8] **Kyul Ko,** Dokyun Son and hyungcheol Shin, “3D TCAD-based Optimization of Channel Diameter considering Interface Trap Variation on GAA Nanowire FET” Korean Conference on Semiconductors, 2016

[9] **Kyul Ko,** Sung-Won Yoo, Hyungwoo Ko and Hyungcheol Shin, “Analysis and verification of TAT GIDL Current Variation Induced by the variation of electric field at Generation-Recombination site and SRH current in MOSFET” 대한전자공학회, 2015.

[10] **Kyul Ko,** Sung-Won Yoo, Hyunseul Lee, Youngsoo Seo, Sangbin Jeon, Hyungwoo Ko, Jeon-Hyun Ok and Hyungcheol Shin, “Analysis of TAT Current Variation Induced by the Slow trap in Silicon” 대한전자공학회, 2015.

[11] Yeaji Yoo, **Kyul Ko**, Myunggon Kang, Jongwook Jeon, and Hyungcheol Shin, “Interplay Between Line Edge Roughness and Interface Traps in Nanoplate VFETs”, Korean Conference on Semiconductors, 2019.

[12] Shinkeun Kim, Dokyun Son, **Kyul Ko**, Myounggon Kang, and Hyungcheol Shin, “Statistical Analysis of NBTI Considering Trap Position in Nanosheet FET” Korean Conference on Semiconductors, 2018.

[13] Dokyun Son, **Kyul Ko,** Changbeom Woo, Myounggon Kang and hyungcheol Shin, “Source-Drain trench based Optimization of 6-T SRAM performance for 3D Stacked Nanowire FET Devices” Korean Conference on Semiconductors, 2017.

[14] Minsoo Kim, **Kyul Ko**, Changbeom Woo, Myounggon Kang, and Hyungcheol Shin, “Intrinsic Chanracteristics and Process Variation Effect of Nanoplate Vertical FET Devices” Int'l Conference on Semiconductor Physics and Devices, 2017.

[15] Changbeom Woo, **Kyul Ko**, Minsoo Kim, Myounggon Kang, and Hyungcheol Shin, “Analysis of RC Delay for Raised Source Multi-channel Vertical Nanoplate FET” Int'l Conference on Semiconductor Physics and Devices, 2017

[16] Dokyun Son, **Kyul Ko**, Myounggon Kang, Sunhom Steve Paak, and hyungcheol Shin, “Impact of Line Edge Roughness (LER) on Single and Three Stacked Nanowire FET in 5 nm node Logic devices” International Vacuum Congress, 2016.

[17] Dokyun Son, **Kyul Ko**, Myounggon Kang and hyungcheol Shin, “Characteristics of Vth variation according to parameters of line edge roughness (LER) in 5 nm node Nanowire FET” NanoKorea, 2016.

[18] Dokyun Son, **Kyul Ko** and hyungcheol Shin, “3D TCAD-based Optimization of Channel radius considering Line edge roughness on GAA Nanowire FET” Korean Conference on Semiconductors, 2016.

[19] Changbeom Woo, **Kyul Ko**, Minsoo Kim, Myounggon Kang and Hyungcheol Shin, “Analysis and Optimization of RC Delay According to Parameter Characteristics in Vertical FET ”International Semiconductor Device Research Symposium, 2016.

[20] Youngsoo Seo, Sung-Won Yoo, Hyoungwoo Ko, Hyunok Jeon, **Kyul Ko**, and Hyungcheol Shin “Trap Type Dependence of Modulation in TAT Gate-Induced Drain Leakage” ITC-CSCC, 2015.

[21] Hyungwoo Ko, Sung-Won Yoo, Youngsoo Seo, Hyunseul Lee, Sangbin Jeon, Hyunok Jeon, **Kyul Ko**, and Hyungcheol Shin, “Electric field variation by single trap considering the relative permittivity variation due to doping concentration” 대한전자공학회, 2015.

[22] Sung-Won Yoo, Hyunseul Lee, Youngsoo Seo, Sangbin Jeon, Jeon-Hyun Ok, **Kyul Ko,** Hyungwoo Ko and Hyungcheol Shin, “Statistical analysis on trap characteristics causing gate-induced drain leakage current random” 대한전자공학회, 2015.

[23] Hyunseul Lee, Sung-Won Yoo, Youngsoo Seo, Sangbin Jeon, Hyungwoo Ko, Jeon-Hyun Ok, **Kyul Ko** and Hyungcheol Shin, “Analysis on the impact of the oxide trap in MOSFET using device simulation” 대한전자공학회, 2015.

[24] Sangbin Jeon, Sung-Won Yoo, Hyunseul Lee, Youngsoo Seo, Hyungwoo Ko, **Kyul Ko**, Jeon-Hyun Ok and Hyungcheol Shin, “New method for extracting Gate Induced Drain Leakage (GIDL) at planar MOSFET using new method” 대한전자공학회, 2015.

[25] Jeon-Hyun Ok, Sung-Won Yoo, Hyunseul Lee, Youngsoo Seo, Sangbin Jeon, Hyungwoo Ko, **Kyul Ko** and Hyungcheol Shin, “Extraction of Distance between Interface Trap and Oxide Trap depending on trap type by consider” 대한전자공학회, 2015

[26] Sangbin Jeon, Sungwon Yoo, Hyoungwoo Ko, **Kyul Ko** and Hyungcheol Shin “A method for extracting parasitic capacitance at planar MOSFET” 대한전자공학회, 2015.

[27] Changbeom Woo, Dokyun Son, **Kyul Ko**, Myounggon Kang and hyungcheol Shin, “Parasitic Capacitance Delay for Vertical Field Effect Transistor”대한전자공학회, 2016.

초 록

본 논문에서는 다양한 공정 변동 요인에 의한 영향을 초소형 GAA FET 소자 및 3차원 NAND Flash Memory 소자에서 정확하게 예측하기 위한 기계 학습 접근법을 제시하였다. 공정 변동성 요인에 의한 영향은 로직 소자와 메모리 소자에서 여러가지 신뢰성 문제의 원인으로 작용하며 특히, 로직 및 메모리 소자의 수율을 결정하는 마진을 감소시켜 정확한 예측 및 제어가 필수적이다.

기계학습 시스템은 크게 비지도적 학습(=Unsupervised Learning), 지도적 학습(=Supervised Learning), 강화 학습(=Reinforcement Learning)의 3가지 계열로 구분된다. 이 중, 소자 특성을 분석하고 변동성 영향 예측을 목적으로 하는 경우 정해진 입출력(Training data) 값에 근거하여 회귀론적 방법으로 예측 모델을 학습시키는 지도적 학습 계열의 기계학습 시스템이 가장 적합한 방법이다. 지도적 학습 계열의 기계학습 시스템은 다양한 변동성 요소에 대한 다각도의 소자 특성을 예측하여야 하기 때문에 다중 노드(=Multi-Node, MN)를 갖는 복잡한 알고리즘(e.g., Artificial neural networks)기반의 다중 입력-다중 출력(=Multi-Input/Multi-Output, MIMO)을 통해 제시되었다.

기계학습 시스템의 초기 단계로 단일 트렌지스터의 변동성 요인을 선행 분석하였다. 초소형 GAA (Gall-All-Around) VFET (Vertical FET) 디바이스의 프로세스 변동 (PV)을 사용하여 주요 전기 매개 변수의 변동을 예측하는 정확하고 효율적인 기계 학습 (ML) 방식을 제시하였다. 제안 된 기계 학습 접근법은 3D 확률론적 TCAD 시뮬레이션과 비교했을 때 동일한

정확도와 우수한 효율성을 보여준다. 인공 신경 네트워크 기반 (ANN) 기계 학습 알고리즘은 MIMO (Multi-input-Multi-Output) 예측을 매우 효과적으로 수행 할 수 있다.

기계 학습 시스템의 발전된 단계로써, 3D NAND 플래시 메모리의 주요 전기 매개 변수의 변화를 예측하는 가변성 인식 기계 학습 시스템을 제안한다. 우리는 최초로 인공 신경 네트워크 (ANN) 알고리즘 기반 ML 시스템의 예측 영향 요인 효과의 정확성, 효율성 및 일반성을 검증하였다. 따라서 다양한 변동 원인으로 인한 장치의 주요 전기적 특성 변화가 동시에 통합적으로 예측된다. 이 알고리즘은 3D 확률론적 TCAD 시뮬레이션을 벤치마킹하여 1 % 미만의 예측 오류율과 80 % 이상의 계산 비용 절감을 보여줍니다. 또한, 층수가 증가함에 따라 다양한 구조 조건을 갖는 3 차원 낸드 플래시 메모리의 동작 특성을 예측함으로써 알고리즘의 일반성을 확인할 수 있다.

주요어 : PV, ML, ANN, GAA, Vertical device, NAND Flash, Prediction

학 번 : 2015-20822