TCAD Augmented Machine Learning for Semiconductor Device Failure Troubleshooting and Reverse Engineering

Y. S. Bankapalli and H. Y. Wong*
Electrical Engineering
San Jose State University
San Jose, CA, USA
*hiuyung.wong@sjsu.edu

***Abstract*— In this paper, we show the possibility of using Technology Computer Aided Design (TCAD) to assist machine learning for semiconductor device failure trouble shooting and device reverse engineering. When TCAD simulation models and parameters are properly chosen and calibrated, large number of devices with random defects and structural characteristics can be generated and simulated. The results can then be used to train machine learning algorithms to predict the defect and structural characteristics of a device with given electrical characteristics (such as IV's and CV's). 1D PIN diode with various layer thicknesses and doping concentrations are used in this study. It is showed that with less than 2000 training samples, by using simple linear regression, one can achieve good prediction of layer thickness and doping of a given IV curve.**

Keywords—Machine Learning, Reverse Engineering, TCAD, Semiconductor Defects

I. Introduction

Semiconductor device failure troubleshooting and device reverse engineering require expensive analyses such as SEM and TEM [1]. Machine learning (ML) has been used widely in the manufacturing process to enable early discovery of defects [2]. However, the authors are not aware of any extensive application of ML to analyze defects based on fabricated device electrical characteristics, such as Current-Voltage (IV) and Capacitance-Voltage (CV) curves, where defects include epitaxial layer thickness and doping level variations. This is probably because, for a matured process with high yield, the number of defective dies is limited, while for nascent process with low yield, the number of dies produced are limited. As result, it is difficult to obtain enough defective IV curves for accurate machine learning.

Using TCAD, in principle, a large number of IV's and CV's can be generated by changing the layer thicknesses (to model epitaxial layer variation) and doping levels (to model doping variation), and by including various defective models (such as trap assisted tunneling at various spatial location). ML can then be used to generate model to accurately correlate IV and CV curves to defect characteristics. Using the trained model, one can rapidly narrow down the possible cause of an abnormal IV or CV curve and, if necessary, perform further failure analysis (e.g. cutting TEM at the most probable failure spot predicted by ML). The same reasoning applies well in device reverse engineering.

In this paper, we demonstrate this idea by studying the relationship between 1-D PIN diode structural defects (epitaxial layer thickness and doping concentration variations) and its forward and reverse IV curves. Various machine learning models are tested. To reduce the number of TCAD simulations, epitaxial layer thickness and doping concentration studies are performed separately.

II. TCAD Simulations

Figure 1 inset shows the structure simulated in which only layer thickness variations are studied. About 2000 1D PIN diode structures are created using SProcess [3] with n+/i/p+ thicknesses being varied independently and uniformly within the range given in Figure 1. Figure 2 shows the scattering plots of n+/i/p+ thicknesses, which are uniform and independent. Sdevice [4] is then used to simulate the IV characteristics. Essential physics models are turned on, including Fermi-Dirac statistic, doping dependent and high field saturation models for carrier mobilities, Schottky-Reed-Hall Recombination (SRH) and non-local Band to Band tunneling (BTBT). 80-bit ExtendedPrecision is used to avoid noisy reversed curves. Poisson, electron and hole continuity equations are solved self-consistently to produce the curves in Fig. 1.

Figure 5 shows the IV's of another 2000 1D PIN diodes simulated with layer concentrations varied independently and uniformly in their logarithmic values. The corresponding inset shows the structure simulated and the variation range. Figure 6 shows that the layer concentrations are independent and their logarithmic values are uniformly distributed.

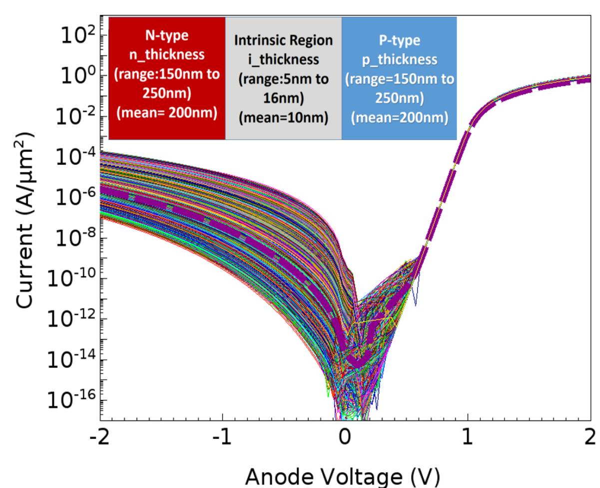

Figure 1: IV's of the 2000 devices (thickness variations only) simulated. The thick pink dash line is the IV of nominal device (200nm/10nm/200nm). Both n+ and p+ concentrations are 10^{20}cm^{-3}. i-layer concentration is 10^{17}cm^{-3}.

978-1-7281-0940-4/19/$31.00 ©2019 IEEE

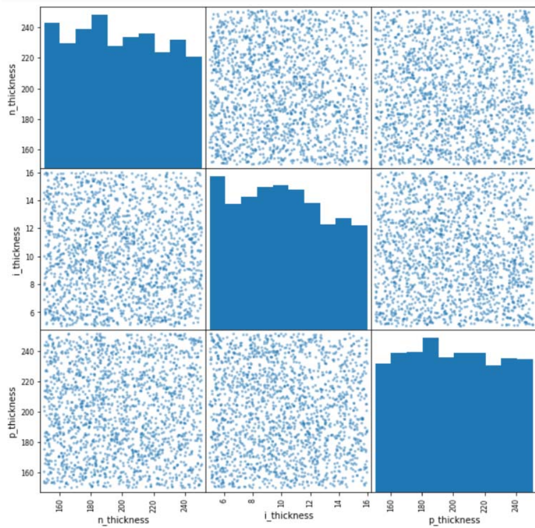

Figure 2: Scattering plot of n+/i/p+ thicknesses showing their frequencies and correlations.

III. Machine Learning for Thicknesses Prediction

Scikit-learn library is used for ML [5]. Four types of algorithms were tested on the layer-thickness data, namely, linear regression (LR), decision tress (DT), random forest (RF) and Multi-Layer Perceptron (MLP) Regressor. 80% of the data (~1600) are used for training and 20% of the data (~400) are used for validation. The input is the IV curve (102 current values for V = -2V to 2V) and the output are i_thickness and n_thickness. Various parts of the IV curve are used for training, namely, I(V=-2V), I(V=-2V to 0V), I(V=0V to 2V) and I(V=-2V to 2V).

The first attempt to train the machine with the raw data was not successful. This is because the current changes orders of magnitude for various thicknesses in reverse bias. As showed in Figure 3, the model fails to predict large i_thickness (prediction is capped at about 12nm) because reverse current (I) is indistinguishable numerically in the raw form for large i_thickness. Moreover, at sufficiently low current level (i.e. when i-thickness is sufficiently large), SRH will dominate and has very weak dependent on the layer thicknesses. If log(I) is used, it gives much better prediction. Therefore, in all trainings for layer thicknesses, log(I) is used.

Table 1 shows the i_thickness and n_thickness prediction Mean Squared Error (MSE) of various machines trained by different data ranges and algorithms. The learning can be summarized as:

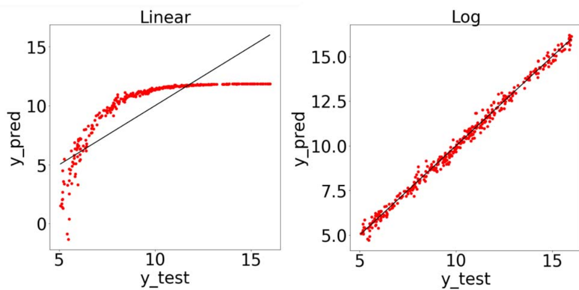

Figure 3: Prediction of i_thickness using linear regression model trained with raw current data (I, left) and processed current data (log(I), right) at V = -2V.

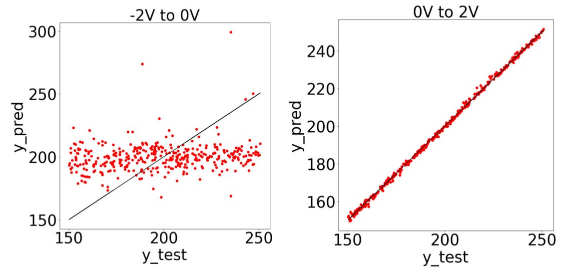

Figure 4: n_thickness prediction by LR machines trained by data from -2V to 0V (left) and data from 0V to 2V (right).

1) DT is a not a suitable algorithm as it often overfits (training MSE = 0, with large prediction MSE)

2) LR performs the best with low training and prediction MSE for both i_thickness and n_thickness

3) MLP performs similar to LR for i_thickness but fails with n_thickness

4) Wide voltage range (-2V to 2V) gives the most accurate results. However, depending on the problem of interest, reduced voltage range gives similar results and simulation time can be substantially reduced. For example, by using the current at -2V, high accuracy of i_thickness can be obtained already because i_thickness influences the BTBT current strongly.

5) It is important to perform the simulation in a regime where the relevant physics is captured. For example, reverse current is insensitive to n_thickness. Therefore, bad result is obtained if data is only available between -2V to 0V. Positive bias simulation is required for n_thickness as forward neutral region potential drop correlates strongly to n_thickness. (Figure 4)

Training of p+ layer thickness gives similar results as the n+ layer thickness and are not shown.

	Data Range Used	MLP	LR	DT	RF
i_thickness	"-2V"	0.06/0.06	0.06/0.06	0.03/0.08	0.01/0.08
	"-2V to 0V"	0.26/0.22	0.04/0.05	0.02/0.06	0.01/0.05
	"0V to 2V"	0.86/0.88	0.09/0.09	0.00/0.29	0.03/0.18
	"-2V to 2V"	0.05/0.05	0.03/0.03	0.01/0.04	0.01/0.04
n-thickness	"-2V"	847/796	846/795	0.00/1741	163/1248
	"-2V to 0V"	1043/984	761/811	0.00/1456	114/751
	"0V to 2V"	514/434	0.96/0.89	0.00/294	22/162
	"-2V to 2V"	473/407	0.80/0.86	86.91/236	22/163

Table 1: Training and prediction Mean Squared Errors (MSE) of i_thickness and n_thickness by machines trained by various data range and algorithms. The numbers are format in "training MSE/ prediction MSE".

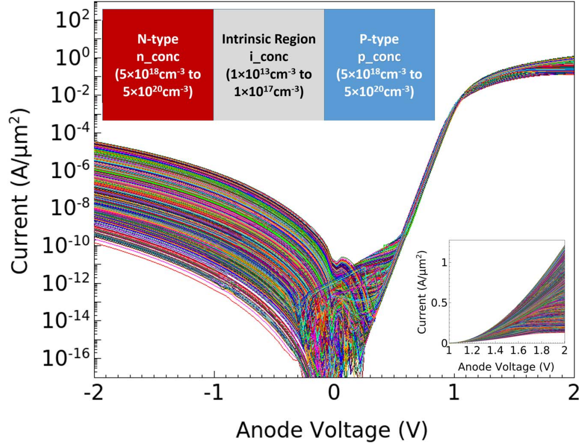

Figure 5: IV's of the 2000 devices simulated with layer concentrations varied. The thicknesses of n+/i/p+ are 200nm/10nm/200nm. The upper left inset shows the structure and the range of variation. The lower inset shows the forward IV's in linear scale.

IV. ML FOR CONCENTRATIONS PREDICTION

Since linear regression shows excellent results in predicting the layer thicknesses of an 1D PIN diode, it is also used to train the model to predict the layer concentrations as the non-linearity is expected to be the similar or less. The structures, IV's and variations are showed in Figures 5 and 6.

The IV distribution of concentration varying diodes is showed in Figure 5 and is very different from that of layer thickness varying one in Figure 1 in the forward region. In Figure 1, the thicknesses, thus the neutral region resistance, vary less than 2 times. But in Figure 5, the concentrations vary by 100 times, which results in large variation of resistance and, thus, forward current. Moreover, from the forward current traces in the inset of Figure 5, one can see that in additional to the magnitude, the shape and curvature vary for different doping concentrations. For example, when the concentration is high, the curvature is positive in the whole region. But when the concentration is low, the curvature changes from positive to negative as voltage increases.

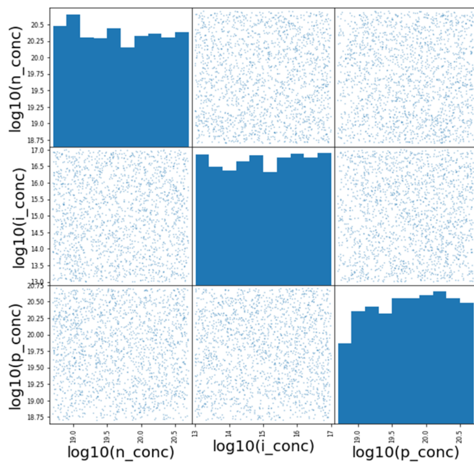

Figure 6: Scattering plot of n+/i/p+ layers concentrations showing their frequencies and correlations used in the study. Note that the logarithmic values of the concentrations are uniformly distributed.

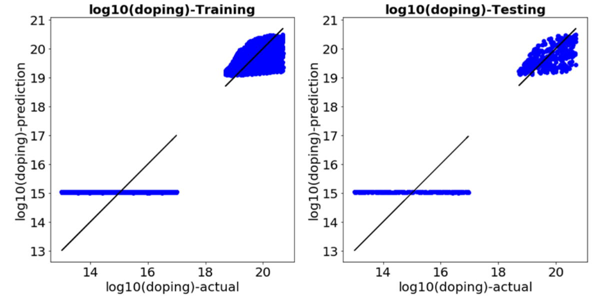

Figure 7: Training and validation of doping concentrations based on current from V = 1.8V to V = 2V.

Therefore, it is expected that the forward region IV will contain sufficient information of the n+ and p+ layer concentrations. Moreover, since it is in forward region, the current values are used directly in the training instead of their logarithmic values being used.

Since n+ and p+ concentrations both have big impact to the IV, multi-variate linear regression is used instead of multi-variation linear regression (which is equivalent to multiple independent linear regression).

Firstly, currents from V = 1.8V to V = 2V are used for training. As shown in Figure 7, although it is expected that doping should have strong impact on the diode current at high forward bias, the training result is very bad. There are two groups of data. The lower concentration one is of the i-layer while the higher concentration one is of the n+ and p+ regions. Since i-layer doping has very small effect on the forward current (due to its small thickness), the model cannot predict any variation in i-conc.

Next, the whole forward current curve is used for training (V = 0V to V = 2V) as showed in Figure 8. The n+ and p+ concentrations can be predicted accurately. It is worth to mention that heavily doped n+ and lightly doped p+ diode (e.g. 5×10^{20}cm^{-3} n+ / 5×10^{18}cm^{-3} p+) is expected to give similar current at V = 2V as the lightly doped n+ and heavily doped p+ diode (e.g. 5×10^{18}cm^{-3} n+ / 5×10^{20}cm^{-3} p+). However, the trained algorithm still can distinguish them clearly. This implies that the asymmetric of n+/p+ doping is captured in the forward IV curves. Indeed, Figure 9 shows that n/p = 5×10^{20}cm^{-3} / 5×10^{18}cm^{-3} and n/p = 5×10^{18} cm^{-3} / 5×10^{20}cm^{-3} give different forward IV shapes. This is probably

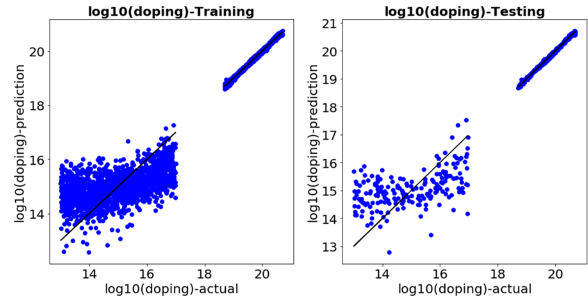

Figure 8: Training and validation of doping concentrations based on current from V = 0V to V = 2V.

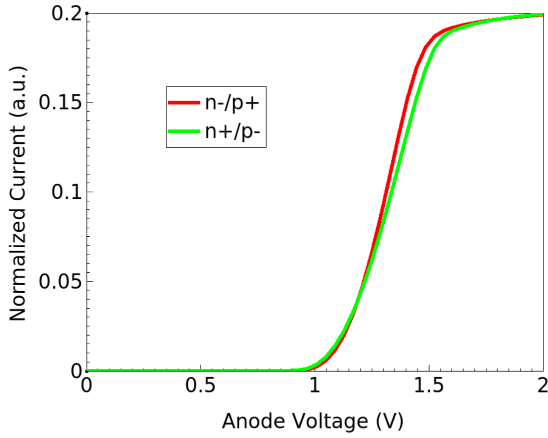

Figure 9: Forward IV of diode with n-/p+ = ($5\times10^{20}cm^{-3}$ /$5\times10^{18}cm^{-3}$) and n+/p- = ($5\times10^{18}cm^{-3}/5\times10^{20}cm^{-3}$). Doping of i-layer is p-type = $10^{17}cm^{-3}$. The current at V = 2V are scaled to be the same.

the reason why the machine can distinguish n and p concentration from each other.

As shown in Figure 8, i-layer concentration still cannot be modeled well even the full forward IV is used in the training. This is because it does not have strong influence on the forward IV due to its small thickness.

Instead of using linear regression with the original 50 input features (i.e. currents at voltage 0V to 2V), second order linear regression is used in which the number of features is expanded to 1326. This gives better fitting in both training and validation. However, it is still not good enough (Figure 10).

Third order linear regression was also tried but it results in overfitting in which it gives perfect fitting to the training model but bad prediction in validation. Therefore, in order to capture the i-layer concentration, more data points are needed.

V. Prospect of 3D TCAD Simulation with ML

The 1D PIN diode has about 300 mesh points. The simulation was performed in Intel Xeon E5-2603 with 1 core used. The total simulation time of each simulation (process and device) is about 90 seconds. As a result, it takes about 2 days to complete the data generation. A typical realistic 3D FinFET IV simulation is between 1 hour to 6 hours (process + device) [6]. If computing farm with thousands of CPU are available, we anticipate similar study can be completed in <1 day for a realistic 3D FinFET structure. To reduce simulation time, one can reduce the number of training data point or/and the range of defect (e.g. i_thickness) variation. Figure 11 shows that even with 50 (or 200) data points, instead of 1600, MLP (or LR) is still very accurate. Moreover, LR can predict accurately 50% wider range of i_thickness than the training data. These make 3D TCAD augmented ML for defect trouble-shooting more feasible.

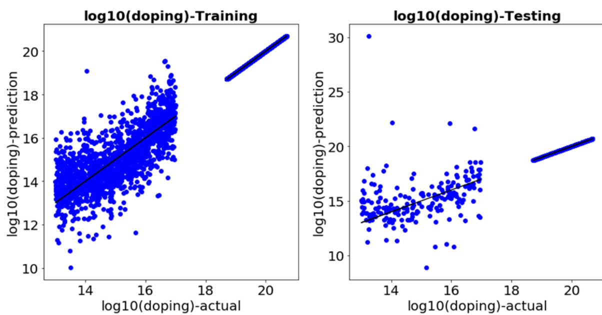

Figure 10: Training and validation of doping concentrations based on current from V = 0V to V = 2V using second order linear regression.

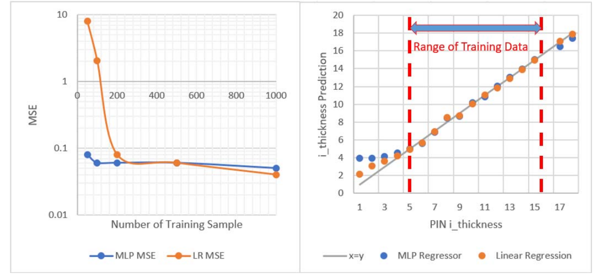

Figure 11: Prediction MSE as a function of the number of training samples (Left). Prediction of i_thickness as a function of PIN i_thickness (Right).

VI. Conclusions

Using PIN diode with various layer thicknesses and concentrations, we demonstrated that TCAD can be used to generate sufficient data to train machine to identify the “defect value” (variation of layer thickness and concentration) rapidly based on IV curves. It is found that 1) data processing before ML is critical to obtain accurate results but the type of preprocessing depends strongly on domain knowledge (e.g. forward and reverse currents require different treatments), 2) linear regression gives the best prediction and is better than Multi-Layer Perceptron (MLP), 3) the model is able to predict structure with thickness out of the range of training data set and 4) the full process (TCAD simulation and ML) can be completed in less than 2 days with 1 cpu core. We anticipate that by using computing farm with thousands of cores, such scheme can be implemented for more realistic 3D simulations. In certain algorithm, only number of training data as low as 50 is needed. Such TCAD augmented ML can expedite defect trouble-shooting and reverse engineering of semiconductor devices.

References

[1] R. Torrance and D. James, "Reverse Engineering in the Semiconductor Industry," 2007 IEEE Custom Integrated Circuits Conference, San Jose, CA, 2007, pp. 429-436. doi: 10.1109/CICC.2007.4405767
[2] G. A. Susto, M. T. and A. Beghi, “Anomaly Detection Approaches for Semiconductor Manufacturing”, Procedia Manufacturing 11 (2017) 2018 – 2024.
[3] Sentaurus™ Process User Guide Version O-2018.06, June 2018.
[4] Sentaurus™ Device User Guide Version O-2018.06, June 2018.
[5] https://scikit-learn.org/stable/
[6] H. Y. Wong, D. Dolgos, L. Smith and R. V. Mickevicius, “Modified Hurkx Band-to-Band-Tunneling Model for Accurate and Robust TCAD Simulations”, submitted to Mircoelectronics Reliabily.