IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 22 October 2016; revised 30 November 2016; accepted 2 December 2016. Date of publication 8 December 2016; date of current version 23 January 2017.

Digital Object Identifier 10.1109/JXCDC.2016.2636161

Physics-Inspired Neural Networks for Efficient Device Compact Modeling

MINGDA LI[1], OZAN İRSOY[2], CLAIRE CARDIE[2], AND HUILI GRACE XING[1]

[1]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850, USA
[2]Department of Computer Science, Cornell University, Ithaca, NY 14860, USA

CORRESPONDING AUTHORS: H. G. XING and M. LI (grace.xing@cornell.edu; ml888@cornell.edu)

This work was supported in part by the Center for Low Energy Systems Technology, one of the six SRC STARnet Centers, in part by MARCO and DARPA, and in part by the National Science Foundation and Air Force Office of Scientific Research EFRI 2-DARE under Grant 1433490.

ABSTRACT We present a novel physics-inspired neural network (Pi-NN) approach for compact modeling. Development of high-quality compact models for devices is a key to connect device science with applications. One recent approach is to treat compact modeling as a regression problem in machine learning. The most common learning algorithm to develop compact models is the multilayer perceptron (MLP) neural network. However, device compact models derived using the MLP neural networks often exhibit unphysical behavior, which is eliminated in the Pi-NN approach proposed in this paper, since the Pi-NN incorporates fundamental device physics. As a result, smooth, accurate, and computationally efficient device models can be learned from discrete data points by using Pi-NN. This paper sheds new light on the future of the neural network compact modeling.

INDEX TERMS Supervised learning, artificial neural networks, semiconductor device modeling, TFETs.

I. INTRODUCTION

DEVICE compact modeling bridges device science to applications, and therefore, it plays a very important role in device research. There are two extremes for device modeling, one is purely physical and the other is purely empirical. Looking at these two extremes, a purely physical modeling method, such as NEMO [1], is computational expensive for use in circuit simulations, and a purely empirical modeling method, such as table lookup model, has limited generalization (extrapolation) ability. Therefore, to find a middle ground between purely physical and purely empirical models, the Electron Design Automation industry, represented by the Compact Model Coalition, chooses to promote physics-based compact models. These use fundamental device physics as the building blocks, then add empirical fitting to modify and merge different analytical physical expressions into smooth functions. However, developing high-quality physics-based compact models is very time-consuming, and therefore often not available for emerging devices. As an alternative, regression with machine learning can be used to model relationships between different variables with certain generalization abilities. Among different regression algorithms, the neural network modeling method has raised a lot of interests [2]–[4] given the fact that it is theoretically capable of arbitrarily accurate approximation to any function and its derivatives [5].

Compared with another widely used data-driven model: table lookup model, the neural network model performs better on the following three aspects.

1) *Scalability:* In order to achieve certain level of accuracy, the table lookup model needs a large amount of data, and the space complexity increases exponentially with increasing dimensions. In contrast, the neural network model is lightweight and scalable.
2) *Generalization:* The table lookup model has poor generalization performance. The polynomial fitting used in the table lookup model often has high out-of-sample errors. In contrast, by using correct learning algorithms, neural network model can be well generalized, which make it more robust against noises.
3) *Smoothness:* An ideal compact model needs to be infinitely differentiable. The table lookup model is not infinitely differentiable due to the nature of polynomial fitting, while using higher order polynomial fitting will improve the smoothness, and it is at the expense of computation efficiency. Therefore, the table lookup

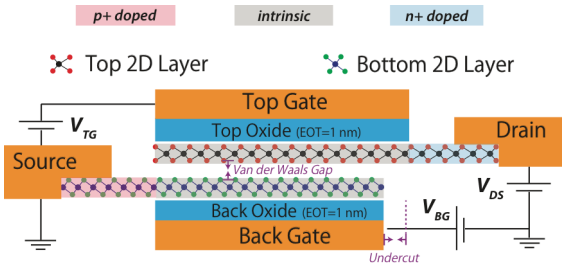

FIGURE 1. Schematic of the example emerging device modeled in this paper. An n-type thin-TFET [7], [8]. Its *I–V* curves are obtained by sweeping the top gate (V_{TG}) with the back gate (V_{BG}) grounded.

model is not possible to be both smooth and computationally efficient. In contrast, the neural network model is guaranteed to be infinitely differentiable.

Previous works [2]–[4] used multilayer perceptron (MLP) neural networks to develop compact models, which are prone to having unphysical behavior [see Fig. 4(e) and (f)]. To eliminate the unphysical behavior, we have developed a novel neural network structure: physics-inspired neural network (Pi-NN), with fundamental device physics embedded. As a result, the Pi-NN can be trained to generate an accurate, smooth, and computational efficient device compact model.

II. THIN-TFET AND TRAINING PROCEDURE

To illustrate the principles of Pi-NN, we develop compact models for the dc I–V curves of a transistor. Physics-based device modeling is typically challenging, because the I–V curves are highly nonlinear and require different analytical physical expressions in different bias windows. Therefore, it is usually difficult to handcraft an infinitely differentiable function from these physical expressions. Since high quality physics-based compact models are yet unavailable for emerging devices, such as tunnel field effect transistors (TFETs) [6], the neural network modeling approach has an added attraction. Here, we used a novel device proposed in our group, a thin-TFET [7] (**2-D** heterojunction interlayer **t**unneling **f**ield **e**ffect **t**ransistor), as an example device for testing the neural network modeling techniques. The schematic device structure of an n-type thin-TFET is shown in Fig. 1. The training data are simulated [7] for the top gate voltage ($V_{\rm TG}$) from 0 to 0.4 V and the drain–source voltage ($V_{\rm DS}$) from -0.1 to 0.4 V with a uniform step of 0.01 V, while the test data are for $V_{\rm TG}$ from 0.005 to 0.405 V and $V_{\rm DS}$ from -0.095 to 0.405 V with a uniform step of 0.01 V. The detailed training procedure is shown in Fig. 2. In the preprocessing step in Fig. 2, a scalar function in the form of $\exp(-a(V_{TG}+b))+1$ is multiplied to the output, which helps improve deep subthreshold modeling. The value of a and b is chosen by following the general rules described below. Since this scalar function is used to improve deep subthreshold modeling, we should choose a and b such as

$$\exp(-a(V_{TG}+b))+1 \begin{cases} \cong 1 & \text{when } V_{TG} > V_{TH} \\ \gg 1 & \text{when } V_{TG} < V_{TH} \end{cases} \quad (1)$$

where V_{TH} is the threshold voltage. Therefore, $|b|$ should be smaller than the threshold voltage and a is approximately the slope of I_D–$V_{\rm TG}$ curves in the deep subthreshold region. The final values of a and b are fine-tuned by trial and error. For example, in this paper, the threshold voltage of thin-TFET is around 100 mV, so b is set to be -50 mV. As for a, the subthreshold swing for $V_{\rm TG} < 50$ mV is around 17 mV/decade, and therefore, a is set to be $1/17 \times 2.3 = 0.135$ mV^{-1} (where 2.3 comes from $\ln(10) \approx 2.3$).

III. MLP NEURAL NETWORK MODELING AND UNPHYSICAL BEHAVIOR

In this section, we use the MLP neural network to generate a compact model for the dc I–V curves of the thin-TFET. The MLP neural network architecture and its well-established learning algorithms are shown in Fig. 3 [9]. After some initial training, we choose to use MLP neural networks with two hidden layers and defined its hyperparameter as (i, j), where i is the number of neurons in the first hidden layer and j is the number of neurons in the second hidden layer. Each neuron uses the hyperbolic tangent function $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ as the activation function. By choosing

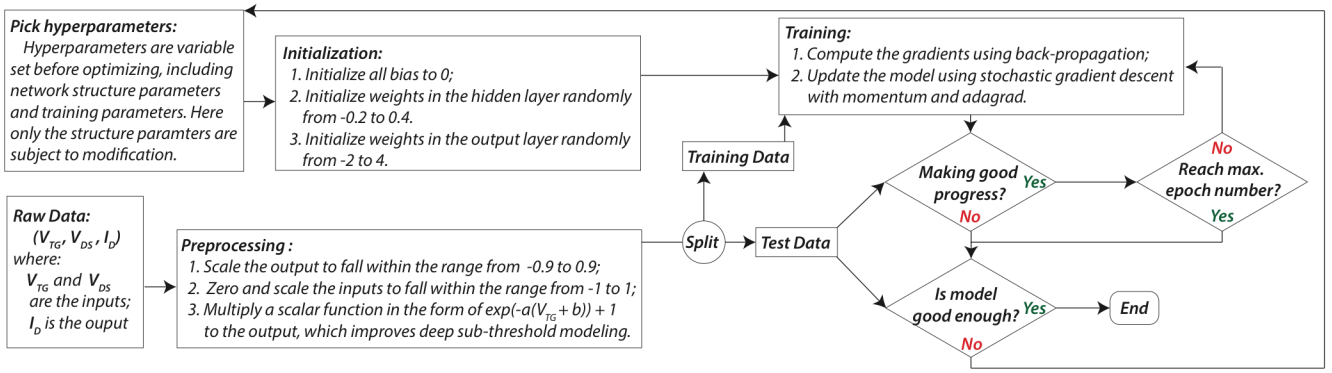

FIGURE 2. Training procedure for artificial neural network device compact modeling.

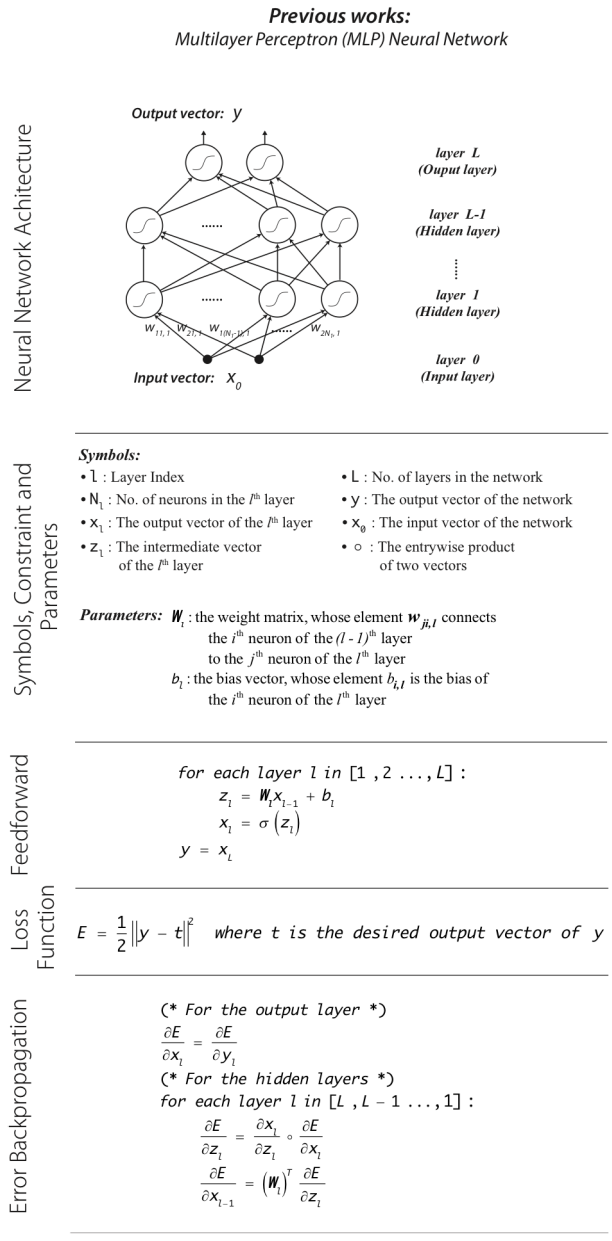

FIGURE 3. Multilayer perception (MLP) neural network model.

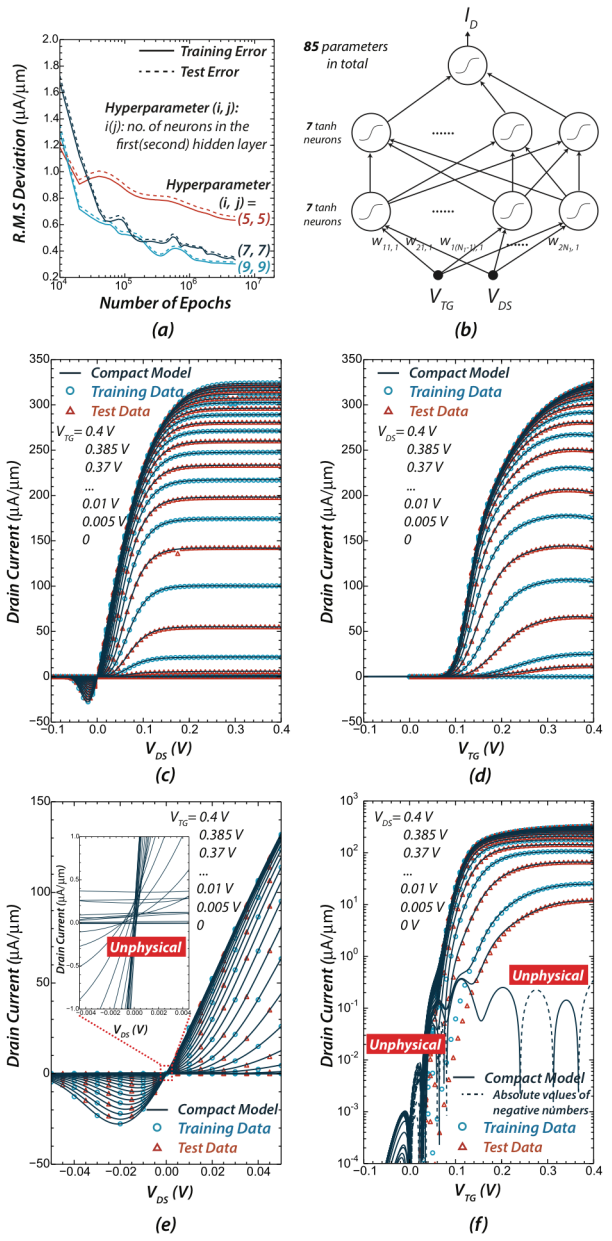

FIGURE 4. Compact model of the n-type thin-TFET derived based on the MLP neural network widely used in previous works [2]–[4]. (a) Training errors and test errors for a variety of hyperparameters. (b) MLP neural network with 7 tanh neurons in the first and second hidden layers. From (c)–(f), the I–V curves generated by the MLP neural network shown in (b) are plotted along with the training data and the test data. (c) I_D versus V_{DS} at different V_{TG} values. (d) I_D versus V_{TG} at different V_{DS} values in linear scale. (e) I_D versus V_{DS} at different V_{TG} values around $V_{DS} = 0$, the embedded plot shows unphysical I_D–V_{DS} relationships around V_{DS} equals 0. (f) I_D versus V_{TG} at different V_{DS} values in semilog scale, unphysical oscillation of I_D around zero appears in the subthreshold region and when $V_{DS} = 0$.

the hyperparameter (i, j) to be (5, 5), (7, 7), and (9, 9), these three MLP neural networks were trained for 5 million epochs. Using the loss function defined in Fig. 3, the root-mean-squared deviations for training data and test data are shown in Fig. 4(a). The test errors are used to evaluate the generalization ability of the model, namely, how the model fit the unseen data. As shown in Fig. 4(a), the test errors stay close to the training errors, which indicated a good generalization. We choose to plot the I–V curves modeled by the MLP neural network with 7 tanh neurons in the first and second hidden layers, as shown in Fig. 4(b), which gives a neural network with 15 neurons and 85 parameters in total. Fig. 4(c)–(f) shows the I–V curves generated by the MLP neural network compact model along with the training data and the test data. Good fitting in the linear scale is achieved for both the I_D–V_{DS} and the I_D–V_{TG} curves. However, if we zoomed-in view the region near $V_{DS} = 0$, I_D is not zero when V_{DS} is zero, indicating that the I_D–V_{DS} relationship is

unphysical around $V_{DS} = 0$ [see Fig. 4(e) (inset)]. Moreover, the I_D–V_{TG} relationship is also unphysical in the subthreshold region [shown in Fig. 4(f)]. The fundamental reason of these unphysical behaviors is that the MLP neural network has no knowledge of the device physics; therefore, the fitting is no longer physical when I_D is very small. In order to eliminate these unphysical behaviors, we have to design a neural network with *a priori* knowledge of the fundamental device physics.

IV. PHYSICS-INSPIRED NEURAL NETWORK DESIGN

First, we note that the inputs V_{DS} and V_{TG} are related to two different physical effects: V_{DS} drives the current through the device while V_{TG} controls the channel potential profile to change the magnitude of the current. Therefore, V_{DS} and V_{TG} should be fed to two different neural networks. According to the fundamental device physics, we know I_D–V_{DS} curves have a linear region at small V_{DS} and a saturation region at large V_{DS}. This behavior is similar to a tanh function. This indicates that V_{DS} should be fed into a neural network with tanh activation functions (tanh *subnet*). To ensure I_D equals zero when V_{DS} equals zero, all the tanh neurons in the tanh *subnet* must have no bias terms. On the other hand, the I_D–V_{TG} curves have an exponential turn-on in the subthreshold region and then become a polynomial in the on region. This is best simulated as a sigmoid function $\text{sig}(x) = 1/(1+e^{-x})$. Therefore, V_{TG} is fed into a neural network with sigmoid activation functions (*sig subnet*). It should be noted that we assumed that gate leakage current is negligible, so V_{TG} would not change the sign of I_D. The final drain current is the entrywise product of the outputs of the tanh *subnet* and the *sig subnet*. This entrywise product reflects the control of V_{TG} on the drain current driven by V_{DS}. In addition, V_{DS} can affect the channel potential profile controlled by V_{TG} due to various nonideal effects, such as the short channel effects. A simple but effective remedy for this is to add weighted connections from each layer in the tanh *subnet* to its corresponding layer in the *sig subnet*. By embedding the above device physics in a neural network structure, we arrive at the Pi-NN. The Pi-NN architecture and its pseudocodes for the feed-forward and error back-propagation algorithms are shown in Fig. 5. This novel neural network is reminiscent of the peephole long–short term memory [10], with the notable difference that the Pi-NN does not propagate through time. After all, the Pi-NN architecture can model the I–V curves of any transistor if two conditions are satisfied: 1) I_D equals zeros if and only if V_{DS} equals zero and 2) V_G does not change the sign of I_D. (i.e., the gate leakage current is negligible).

V. PHYSICS-INSPIRED NEURAL NETWORK MODELING

After initial training, we choose to use Pi-NNs with one hidden layer and define the hyperparameter as (m, n), where m is the number of the tanh neurons in the hidden layer and n is the number of the *sigmoid* neurons in the same hidden layer. The test errors stay close to the training errors, as shown in Fig. 6(a), which indicates good generalization. The model

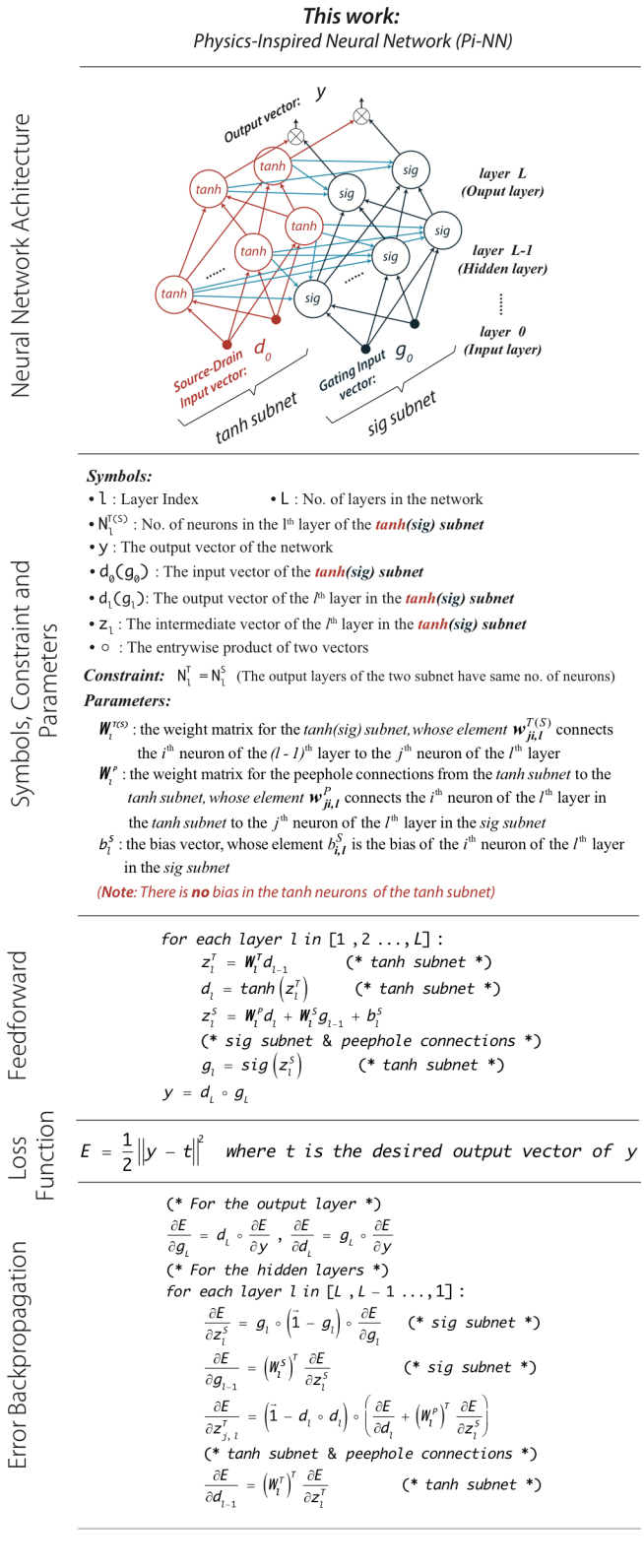

FIGURE 5. Pi-NN model. *Source code available at https://github.com/Oscarlight/Pi-NN.*

complexity is gradually increased from the hyperparameter (2, 2) to (3, 4). From Fig. 6(a), the model with the hyperparameter (2, 3) is the simplest model with converging

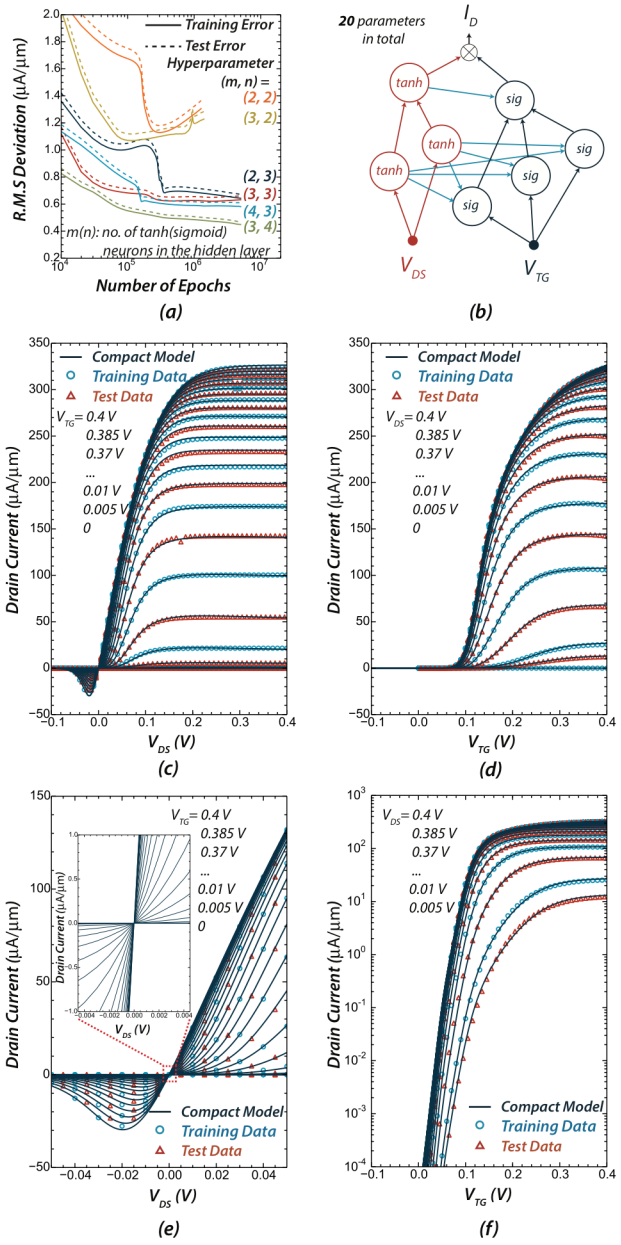

FIGURE 6. Compact model of the n-type thin-TFET derived based on the Pi-NN developed in this paper. (a) Training errors and test errors for a variety of hyperparameters. (b) Pi-NN model with 2 tanh neurons and 3 sigmoid neurons in the hidden layer. From (c)–(f), the *I–V* curves generated by the Pi-NN model shown in (b) are plotted along with the training data and the test data. (c) I_D versus V_{DS} at different V_{TG} values. (d) I_D versus V_{TG} at different V_{DS} values in linear scale. (e) I_D versus V_{DS} at different V_{TG} values around $V_{DS} = 0$, the embedded plot shows well-behaved I_D–V_{DS} relationship around $V_{DS} = 0$. (f) I_D versus V_{TG} at different V_{DS} values in semilog scale, good fitting is achieved in the subthreshold region. All the unphysical behaviors of the MLP neural network are eliminated, and the size of the neural network is largely reduced.

training and test error. More complex models can achieve smaller training and test error but the improvement is not significant enough to justify the increased complexity. Balancing between model complexity and accuracy, we choose the model with the hyperparameter (2, 3), as shown in Fig. 6(b), which give a small Pi-NN model with only 7 neurons and 20 parameters in total. Excellent modeling is demonstrated in both the on region [shown in Fig. 6(c) and (d)] and the subthreshold region [shown in Fig. 6(f)]. The I_D–V_{DS} relationship around V_{DS} equals zero is shown in Fig. 6(e). All the unphysical behaviors that appeared in the MLP neural network model have been eliminated. Moreover, thanks to the embedded device physics, the Pi-NN requires much less parameters than the MLP neural network, which results in a smaller, more efficient compact model.

VI. CONCLUSION

Motivated by the need of high-quality compact models for emerging devices, we have proposed a novel neural network: Pi-NN, for compact modeling. With fundamental device physics incorporated, the Pi-NN method can produce accurate, smooth, and computational efficient transistor models with good generalization ability. Thin-TFET is presented as an example to illustrate the capabilities of Pi-NN. A relatively small compact model is achieved with excellent fitting in both the on and the subthreshold regions of the thin-TFET. The charge–voltage, Q–V, relationships in a device are highly desirable for circuit design. It is possible to construct Q–V relations from the device C–V data (not shown here). However, since the sign of the terminal charge density is dependent on both V_{TG} and V_{DS}, the Pi-NN architecture cannot be directly applied for modeling Q–V relations. The walk-around is to connect V_{TG} and V_{DS} to both the tanh subnet and the *sig* subnet in the Pi-NN, and add the bias terms in the tanh neurons. This modified Pi-NN is compatible with the adjoint neural network method for constructing Q–V relation from C–V measurements [2], [11]. However, this modified Pi-NN architecture has no apparent advantage over the MLP architecture for Q–V modeling. Future work will focus on how to better integrate Q–V modeling into the Pi-NN framework. Finally, the Pi-NN approach is readily implementable in commercial measurement and modeling systems.

REFERENCES

[1] J. Sellier *et al.*, "NEMO5, a parallel, multiscale, multiphysics nanoelectronics modeling tool," in *Proc. Int. Conf. Simulation Semiconductor Process. Devices (SISPAD)*, Denver, CO, USA, 2012.

[2] X. Jianjun, and D. E. Root, "Advances in artificial neural network models of active devices," in *Proc. IEEE MTT-S Int. Conf. Numer. Electromagn. Multiphys. Modeling Optim. (NEMO)*, Aug. 2015, pp. 1–3.

[3] H. B. Hammouda, M. Mhiri, Z. Gafsi, and K. Besbes, "Neural-based models of semiconductor devices for SPICE simulator," *Amer. J. Appl. Sci.*, vol. 5, no. 4, pp. 785–791, 2008.

[4] W. Fang, and Q.-J. Zhang, "Knowledge-based neural models for microwave design," *IEEE Trans. Microw. Theory Techn.*, vol. 45, no. 12, pp. 2333–2343, Dec. 1997.

[5] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, no. 2, pp. 251–257, 1991.

[6] A. C. Seabaugh and Q. Zhang, "Low-voltage tunnel transistors for beyond CMOS logic," *Proc. IEEE*, vol. 98, no. 12, pp. 2095–2110, Dec. 2010.

[7] M. Oscar, D. Esseni, J. J. Nahas, D. Jena, and H. G. Xing, "Two-dimensional heterojunction interlayer tunneling field effect transistors (Thin-TFETs)," *IEEE J. Electron Devices Soc.*, vol. 3, no. 3, pp. 200–207, May 2015.
[8] M. Oscar, D. Esseni, G. Snider, D. Jena, and H. G. Xing, "Single particle transport in two-dimensional heterojunction interlayer tunneling field effect transistor," *J. Appl. Phys.*, vol. 115, no. 7, p. 074508, 2014.
[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1988.
[10] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, pp. 115–143, Aug. 2002.
[11] J. Xu, M. C. E. Yagoub, R. Ding, and Q. J. Zhang, "Exact adjoint sensitivity analysis for neural-based microwave modeling and design," *IEEE Trans. Microw. Theory Techn.*, vol. 51, no. 1, pp. 226–237, Jan. 2003.

MINGDA LI received the B.S. degree in microelectronics from Fudan University, Shanghai, China, in 2012, and the M.S. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 2014. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Cornell University, Ithaca, NY, USA.

OZAN İRSOY received the B.A. degree in mathematics and the B.Sc. degree in computer engineering from Bogazici University, Istanbul, Turkey, in 2012. He is currently pursuing the Ph.D. degree in computer science with Cornell University, Ithaca, NY, USA.

CLAIRE CARDIE received the B.S. degree from Yale University, New Haven, CT, USA, in 1982, and the M.S. and Ph.D. degrees from the University of Massachusetts, Amherst, MA, USA, in 1989 and 1994, respectively, all in computer science.

She is currently a Professor with the Department of Computer Science and the Department of Information Science, Cornell University, Ithaca, NY, USA.

HUILI GRACE XING (S'01–M'03–SM'14) received the B.S. degree in physics from Peking University, Beijing, China, in 1996, the M.S. degree in material science from Lehigh University, Bethlehem, PA, USA, in 1998, and the Ph.D. degree in electrical engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2003.

She is currently a Professor with the School of Electrical & Computer Engineering and the Department of Materials Science and Engineering, Cornell University, Ithaca, NY, USA.