Variation-Resilient FeFET-Based In-Memory Computing Leveraging Probabilistic Deep Learning

Bibhas Mannae, Arnob SahaG, Zhouhang Jiange, *Graduate Student Member, IEEE,* Kai **NiG,** *Member, IEEE,* and Abhronil SenguptaG, *Senior Member, IEEE*

***Abstract*- Reliability issues stemming from device level nonidealities of nonvolatile emerging technologies like ferroelectric field-effect transistors (FeFETs), especially at scaled dimensions, cause substantial degradation in the accuracy of in-memory crossbar-based AI systems. In this work, we present a variation-aware design technique to characterize the device level variations and to mitigate their impact on hardware accuracy employing a Bayesian neural network (BNN) approach. An effective conductance variation model is derived from the experimental measurements of cycle-to-cycle (C2C) and device-to-device (D2D) variations performed on FeFET devices fabricated using 28 nm high-k metal gate technology. The variations were found to be a function of different conductance states within the given programming range, which sharply contrasts earlier efforts where a fixed variation dispersion was considered for all conductance values. Such variation characteristics formulated for three different device sizes at different read voltages were provided as prior variation information to the BNN to yield a more exact and reliable inference. Near-ideal accuracy for shallow networks (MLPS and LeNet models) on the MNIST dataset and limited accuracy decline by ~3.8%-16.1% for deeper AlexNet models on CIFAR10 dataset under a wide range of variations corresponding to different device sizes and read voltages, demonstrates the efficacy of our proposed device-algorithm co-design technique.**

***Index* Terms-Bayesian Neural Network (BNN), device-algorithm co-design, ferroelectric field-effect transistor (FeFET) crossbar, in-memory computing, variation-aware design.**

Manuscript received 24 December 2023; revised 20 February 2024; accepted 12 March 2024. Date of publication 25 March 2024; date of current version 24 April 2024. This material is based upon work supported by the National Science Foundation under Grant CNS 2137259 - Center for Advanced Electronics through Machine Learning (CAEML) and its industry members. The characterization of FeFET device nonidealities was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Energy Frontier Research Centers program under Grant DE-SC0021118 and by the National Science Foundation under Grant 2347024. The review of this article was arranged by Editor P.-Y. Du. *(Correspondingauthor: Abhronil Sengupta)*

Bibhas Manna, Arnob Saha, and Abhronil Sengupta are with the School of Electrical Engineering and Computer Science, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: sengupta@psu.edu).

Zhouhang Jiang and Kai Ni are with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA.

Color versions of one or more figures in this article are available at https://doi.org.110.1109/TED.2024.3378223.

Digital Object Identifier 10.1109/TED.2024.3378223

I. INTRODUCTION

EMERGING nonvolatile memories capable of performing simultaneous compute and storage functionaUties show great promise for the hardware acceleration of deep neural networks [1], [2]. The data-intensive and complex vector-matrix multiplication operations required in neural networks can be realized on-chip learning by harnessing the inherent physical attributes of the memory devices arranged in an array fashion-resulting in "in-memory computing." Among different potential memory candidates such as resistive random access memory **(RRAM),** phase-change memory (PCM), and magnetic devices, hafnia-based ferroelectric field-effect transistor (FeFET) has lately earned great interest due to its CMOS compatibility, low energy operation, multilevel programming capability with wider dynamic range, decoupled read-write operation, easy array-level integration, among others [3], [4], [5], [6]. The voltage-driven partial polarization switching in the ferroelectric (FE) layer of FeFET promotes gradual tuning of channel conductivity, mimicking analog synaptic weight update behavior. However, process variation-induced stochastic variabilities stemming primarily from the polycrystalline FE and their pronounced effect with device scaling poses a serious challenge to accomplishing reliable computing using FeFET crossbars. The device-level nonidealities with read-write fluctuations cause the stored weight (i.e., programed conductance) to deviate significantly from the expected trained value, resulting in drastic accuracy degradation of the neural network at the hardware level. Thus, addressing device-level reUability and proposing practical solutions to combat their consequences are crucial to designing variation-tolerant FeFET-based neuromorphic computing.

Prior efforts in this direction mostly adopt either expensive retraining or repeated evaluation-remapping methods demanding nontrivial design overhead [7], [8], [9]. Some works incorporate generalized noise models in the network weights at the algorithm level and attempt to compensate for their effects through iterative training but are unable to perform the learning task jointly with robustness optimization [10], [11]. Bayesian inference-based approach on memristor-crossbar-based systems considering device nonidealities and stuck-at-faults has been proposed recently to achieve robust computing [12], [13].

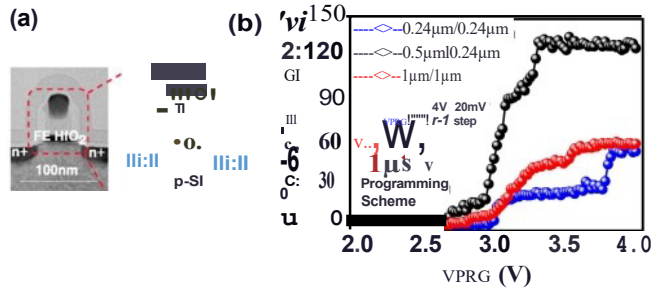

Fig. 1. (a) TEM cross-section and schematic representation of FeFET fabricated on 28 nm HKMG node with doped HfO2 serving as the FE. (b) Conductance-programming voltage characteristics of FeFET for three different device dimensions at read voltage, VRead, of 1.2 V.

However, the proposal is formulated based on a parameterized canonical form of variation derived from a more generalized and hypothetical device model and therefore does not reflect realistic interplay of variations with device dimensions and operating voltage conditions. This work seeks to bridge this critical gap by underscoring the strong need to consider such hardware-software co-design effects in algorithm design supported by extensive experimental variation characterization on industry-scale FeFETs and subsequent performance evaluation on standard machine learning benchmark suites. An empirical conductance variation model derived at the individual device level is coupled to the probabilistic learning-based uncertainty optimizer at the algorithm level to abate the effect of hardware nonidealities on recognition performance. The work possesses the following distinctive novelties.

1) Formulation of a realistic conductance variation model capturing relationship of device nonidealities with scaling and operating voltage conditions through direct experimental characterization of scaled FeFET devices fabricated on industry standard process technology.
2) Design of a practically feasible variation-aware framework by incorporating device size-dependent unique variation characteristics against various programming voltages into the Bayesian neural network (BNN) training framework to deliver robust and stable inference.

The rest of the article is organized as follows: Section II explains the scheme for experimentally characterizing the channel conductance of FeFETs having different gate dimensions. Section III presents design of the proposed Bayesian training framework incorporating device level conductance variation information. The section provides necessary background on the operation of **BNN,** explores the impact of device scaling on channel conductance properties through simulations and experiments, derives an effective device-specific nonideality model, and formulates the prior for BNN. Section IV evaluates performance of the proposed variation-aware framework and discusses its efficacy to mitigate the impact of device-level nonidealities on computing performance. Finally, we outline the key conclusions of our work in Section V.

II. EXPERIMENTAL CHARACTERIZATION OF FEFET

Experimental measurements have been performed on FeFET fabricated using an industrial 28 nm high-k metal gate (HKMG) technology node [14]. Three identical FeFET devices of different gate sizes: W(width)/L(length) = 1/1 $\mu,m,$ 0.5/0.24 µ,m, and 0.24/0.24 μ,m have been considered for the present study. The schematic device structure and corresponding cross-sectional TEM image are shown in Fig. !(a) [14], [15]. Each device comprises a vertical metal-FE-insulator-silicon (MFIS) stack with 8 nm thick doped HfO2 and ~1 nm thick high-quality SiO_2 functioning as FE and insulator layers, respectively. To measure the modulation of channel conductivity in response to polarization switching, each device was subjected to a gate voltage with a pulsing scheme, as illustrated in Fig. 1(b). A reset pulse (VRST) of -4 V preceded every programming pulse (VPRG) to switch all the domains to the negative polarization state, thus resetting the device to the initial lowest conductance state every time. A positive programming pulse of progressively growing amplitude (2-4 V with a step of 20 mV) was employed to access all possible intermediate conductance states of the device. The read-out of the programed state was accomplished immediately after each write operation by applying a ramp gate voltage (VRead)- The drain terminal was held at 0 V during the reset and programming operations, and was switched to 50 mV during the read operation.

III. DESIGN OF VARIATION-ROBUST BAYESIAN TRAINING FRAMEWORK

A. Preliminaries

The proposed algorithmic framework leverages the intrinsic property of BNNs to produce accurate and robust inference under weight fluctuations by incorporating the probability distributions associated with variation data obtained through extensive characterization of the FeFET devices considering effects of device sizing, read noise, among others. The jth weight, w_j, of a network is assumed to be mapped to the conductance state, g_j, within the FeFET programming range present at the jth cross-point, following the relationship [16]:

$$g_j = \left(\frac{g_{mx} - g_{mn}}{w_{mx} - w_{mn}}\right)\left[|w_j| - w_{mx}\right] + g_{mx} \qquad (1)$$

where $8mx$, $8mn$, and Wmx, Wmn are the maximum and minimum values of the conductance and corresponding weight, respectively. As the programed conductance is variational, it is highly appropriate to treat each network weight as a distribution rather than having a specific value, which is fortunately the inherent behavior of BNNs. For a given dataset, D, the BNN accepts a priori, $P(w)$, on variatfon characteristics of noisy weights to find posterior weight distributions following the Bayes rule: $P(w|D) = P(w)P(D|w)/P(D)$. As the true weight posterior, $P(w|D)$, is computationally intractable, stochastic variational inference scheme is applied to approximate $P(w|D)$ with a distribution, $q(w|0)$, that minimizes the Kullback-Leibler (KL) divergence with true Bayesian posterior [17], [18]. The corresponding objective function, 71(D, 0), for optimization becomes

$$71(D, 0) = -Eq(wlll)[\log P(D|w)] + KL[q(wl0)11P(w)). \qquad (2)$$

The approximated posterior distribution of weights, *q(wl0),* in BNNs is learned iteratively through the "Bayes by Backprop" method to enforce that the posterior follows the device variation characteristics supplied as prior information, *P(w),* to the framework [I9]. Considering Gaussian distribution, variational parameter, *0(μ,q, aq),* foreach weight posterior of the BNN is updated by descending along the gradients of the objective function. A reparameterization trick is useful to obtain more efficient gradient estimation for *0,* enabling training iteration to be compatible with standard backpropagation [19]. The data-dependent first component of (2) represents likelihood cost, which is the standard loss function averaged over multiple single network models derived by sampling the posterior weight distribution. On the other hand, the prior-dependent second component represents the KL divergence loss, which computes the degree of dissimilarity between the prior and posterior distributions and accounts for ensuring robustness to the optimization problem. Hence, the framework is capable of simultaneously handling the goal of maximizing accuracy and minimizing reliability induced errors by driving the posterior to follow the prior through the backpropagation method. Upon successful completion of training, the mean values, *μ,q,* of the optimized posterior distributions are regarded as the fully trained weights (mapped to the FeFET conductance at the hardware level) for inference evaluation.

B. Conductance Characteristics of Scaled FeFET

We started our analysis by measuring the conductance-programming voltage characteristics of FeFET for three different gate areas, as presented in Fig. l(b). A more gradual transition of channel conductance with VPRG was observed for the $W/L = 1/1\ \mu,m$ device. The gradual switching with a continuum of states reflects a broader distribution of coercive voltages across a large number of domains (tiny switchable units) in the FE layer such that polarization flipping is possible for a subset of domains at almost every incremental VPRG· However, as the gate area shrinks, the number of domains **in** the FE layer reduces proportionally, and the nonhomogeneity and randomness to the coercive field distribution becomes more pronounced. This certainly introduces nonlinearity to the conductance programming profile with reduction in number of states, as evident from Fig. l(b) for the $W / L = 0.50.24\ \mu,m$ and $0.24/0.24\ \mu,m$ device sizes.

The stochastic polarization switching of individual domains in the FE layer and the process variations involved in device fabrication introduce obvious cycle-to-cycle (C2C) and device-to-device (D2D) variation effects on the conductance, especially for scaled devices. Fig. 2(a) and (b) illustrates the mean and standard deviation of experimentally measured C2C and D2D variations as filled error plots at a VRead of 1.2 V. The intradevice variations measured over 50 cycles indicates that an appreciable amount of C2C variation is present for all devices and is almost insensitive to the device size. On the other hand, the interdevice deviations calculated over three devices for each size implies that D2D variations increases

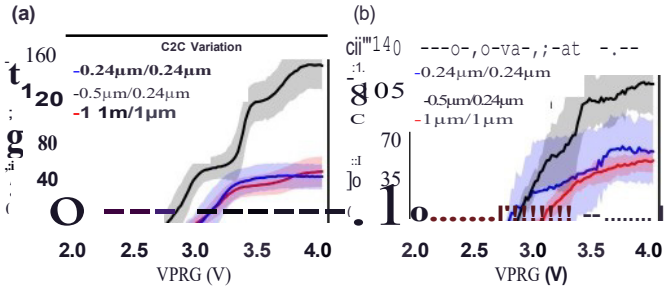

Fig. 2. Filled error plot showing mean (solid line) and associated standard deviation (broadening) of conductance states for (a) C2C variations measured over 50 consecutive programming cycles for each amplitude of VPRG corresponding to a single device of each device size and (b) D2D variations recorded over three devices of the same size by running a singleprogramming pulse for each amplitude of VPRG at VRead of 1.2 V.

drastically with device scaling and dominates over C2C variation for highly scaled devices. The observation agrees well with earlier reports on scaled FeFETs [20], [21]. The higher D2D variation in smaller devices is primarily attributed to reduced domain number, increased in-homogeneity in the domain distributions, more randomness in the distribution of FE and dielectric phases in the FE layer, among others [5], [22], [23], [24], [25]. Although the degree of D2D variations may differ if the experimental characterization is performed over larger number of devices, their dependency on device scaling is expected to remain unchanged.

Furthermore, we substantiated our observation on D2D variations computed over limited experimental data using a well-established Monte Carlo based simulation model [26], [27]. The model considers the poly-crystalline FE layer as an ensemble of multiple uncorrelated domains randomly initialized to either of the two stable polarization states. The switching between states for a domain at any time step, *!1t,* is associated with a finite probability, *Psw,;,* which under the influence of temporally varying electric field, *EFE(t),* can be expressed as

$$\text{Psw,;} = 1 - \exp[\text{h;(t)-6-}\ h;(t + 11t)'8] \tag{3}$$

where /J is the shape parameter of the probability distribution. The history parameter, *h;(t),* which is responsible for accumulating instantaneous stimuli to the ith domain over time can be computed as

$$h;(t) = \int_{t_0}^{t} dt' \frac{}{\text{-Csw,;(EFE(t), } Ea,i)} \tag{4}$$

The domain switching time constant, rsw,;, can be formulated following the nucleation limited switching model [26]. The parameter, *h; (t),* captures polarization accumulation effects and increases over time until the domain reverses its state. The polarization of the entire film at any time is estimated as a summation over all the individual domains. The time-dependent polarization dynamics of the film is next coupled to the conventional charge-voltage equation of the n-channel FET to solve for the channel conductance of the FeFET self-consistently. The values of the parameters used for the device simulation are the same as noted in

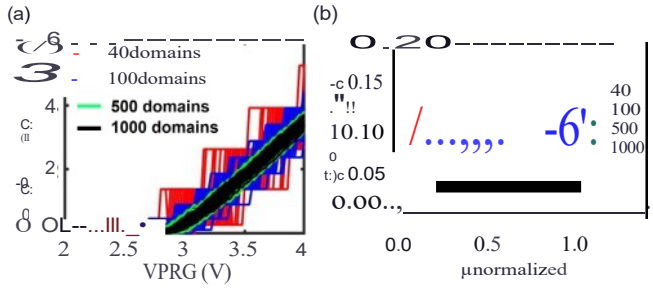

Fig. 3. (a) Simulation results showing D2D variations computed over 200 devices for different number of domains in the FElayer. (b) Standard deviation plotted against mean of the D2D variations after normalizing the conductance programming data to a maximum value of unity.

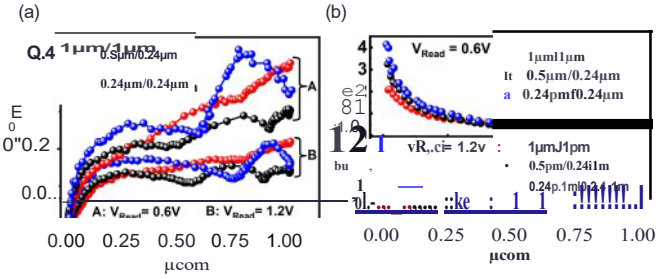

Fig. 4. (a) Standard deviation, <Tcom, as a function of mean, µ00m, of the variation in FeFET programming combining both C2C and D2D measurement data. (b) Severity of variations, <Tcomlµcom, has been plotted against different values of mean,om, at VRead of 0.6 and 1.2 V.

prior work [28], [29]. For the simulation of D2D variations, an activation field value is sampled randomly over a normal distribution of *Ea* for each of the domains in 200 identical devices [21]. The activation field affects the switching probability, *Psw,;,* of the individual domains through the history parameter, *h;(t)* [see (4)]. Since the distributions of *Ea* are not identical across 200 devices containing the same number of domains, the partial polarization switching dynamics of the FE layer is expected to be different from one device to another. The simulated D2D variation, as shown in Fig. 3(a), demonstrates that the variation increases greatly with the decrease in domain number (i.e., with the down-scaling of the device area), thereby in agreement with our experimental findings.

C. Device Variability Modeling

To quantify the amount of variation involved in the conductance programming process and model its dependence on device sizing, an effective variation parameter combining both the spatial (02D) and temporal (C2C) effects was derived based on experimental data. The mean, Jl,com, and standard deviation, D'com, of the combined variation effects was estimated by averaging C2C variations over multiple devices. Fig. 4(a) illustrates such a combined variation profile within the entire programming range for different device sizes at two read voltages of 0.6 and 1.2 V. The exponential rise of D'com for initial smaller values of Jl,com is primarily due to the greater degree of randomness associated with bias-dependent domain polarization switching at relatively weaker VPRG. As the strength of VPRG increases, domains switch more deterministically causing asymptotic decay of variation for higher Jl,com. However, the presence of sharp kinks in the D'com profile can be identified for scaled devices. These sharp transitions correspond to the nonuniform coercive field distribution in the FE film, causing abruptly varying gradient in conductance switching, as can be understood from the simulation data provided in Fig. 3(a) and (b). Though the magnitude of variations reduces with increasing VRead, their nature remains almost insensitive and is characteristic to the device structure (i.e., dispersion of domains in the FE layer). Hence, proper understanding and extraction of device-dependent variation characteristics is extremely important for investigating the impact of device nonidealities at the crossbar array level.

Next, we derive an empirical model equation that best fits the variation characteristics shown in Fig. 4(a) employing a higher order polynomial function. The generalized fitted variation equation as a function of the mean programed conductance state can be expressed as

$$D'com' = \sum_{j=(J}^{n} L C;\mu,::C,m' \qquad (5)$$

where, *Jl,com'* and D'com' are the fitted equivalent of *Jl,com* and D'com, respectively. The coefficients in (5) can be derived to minimize the approximation error. For instance, the parameters for 1/1 *µ,m* device at VRead of 1.2 V have been extracted as: (Co = 0.0258), C1 = 0.788 C2 = -0.0214, and (C3 = 2.1 x 10-4). The relative variation (acorn/Jl,com) plotted in Fig. 4(b) suggests that a significant fraction of the available conductance states undergoes a remarkably high amount of variation and severity increases at lower VRead.

D. Prior Formulation for BNN Training

Simple prior formulation based on a generic variation model utilized in prior works cannot account for hardware-specific dependencies of the variation effects on device scaling, read noise, presence of sharp transitions in variation spectra of scaled devices, among others. Hence, there is an obvious need for reformulating the prior. Considering no correlation among the variations of neighboring devices, we reformulate the prior, *P(w),* employing a univariate Gaussian distribution. While the mean of the prior, *Jl,p,* for each weight follows the mean of the respective posterior, *Jl,q,* at any iteration, the broadening parameter, *ITp,* is estimated from the relative variation [as provided in (5)] experienced by the weight equivalent conductance, D'com'. Such prior formulation approach enables us to efficiently encode the exact variation structural characteristics in the probability distribution of the network weights. This sharply contrasts prior works on BNN, where a Gaussian model with a fixed amount of variation was considered as the prior for all the weights [12], [13].

IV. PERFORMANCE EVALUATION OF PROPOSED FRAMEWORK

The performance of the proposed framework was evaluated for three different neural network architectures [30]: five layered MLP5, LeNet, and AlexNet on MNIST [31]

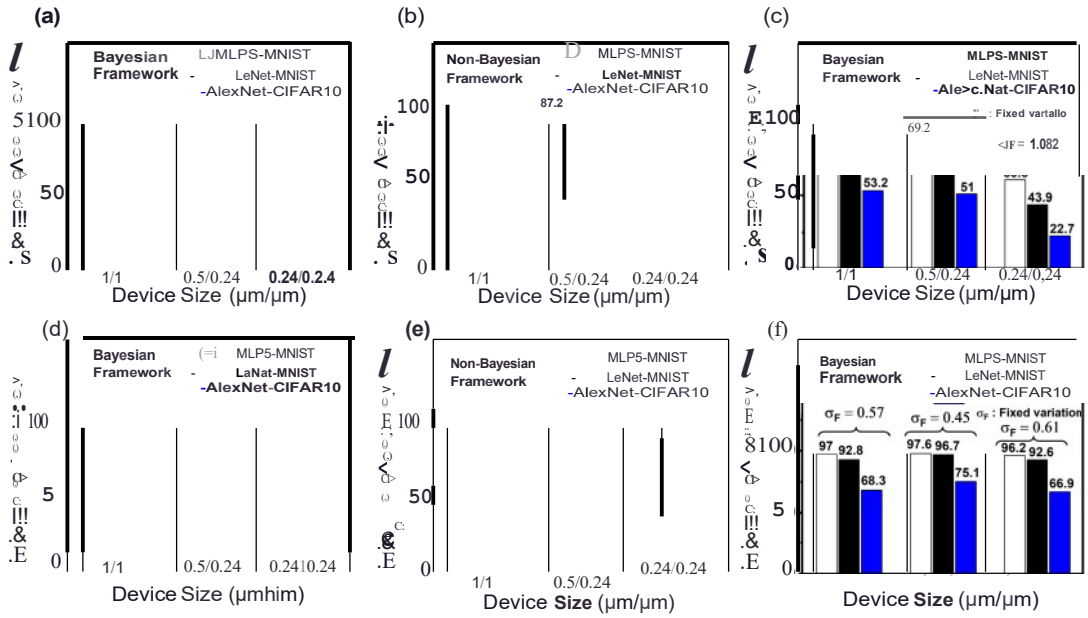

Fig. 5. Bar-chart comparison of inference accuracy for different network models under variations corresponding to different device sizes [following (5)] at VRead of 0.6 V, employing (a) proposed Bayesian and (b) non-Bayesian frameworks. (c) Inference accuracy of network models trained under Bayesian framework but all network weights are subjected to a fixed amount of variation, *aF*, irrespective of programed conductance state. (d)-(f) Inference performance results for different network models evaluated at VRead of 1.2 V applying the same respective schemes as in (a)-(c). The inference outputs in the Bayesian frameworks have been derived by injecting noises [see (5)] to the trained meanweights andaveraging over five runs.

and CIFAR IO [32] datasets. The algorithm for Bayesian learning was developed following the "Bayes by Backprop" method [19] and was implemented in PyTorch.[1] The local reparameterization trick has been employed to reduce computational overhead by translating expensive sampling operation over noisy distributions from high-dimensional weight space to the lower dimensional activation level [33]. The layer-wise KL Joss between prior and posterior *(P(w) ~ N(µ,p, O'p); q(w) ~N(µ,q,O'q))* has been calculated following [34]: KL(P(w)jjq(w)) = *Jog(O'q/O'p)* + (O';/m;) + *(µ,p - µ,q)2*/20'J - 1/2. l0% of KL Joss was added to the standard likefjhood Joss to derive total Joss of the network at every iteration. The network has been trained using Adam optimizer with an efficient learning-rate scheduler and input batch size of 128. The ideal software-based inference accuracies of MLP5-MNIST, LeNet-MNIST, and AlexNet-CIFARlO (architecture-dataset format) without any variations are 98%, 99.1%, and 85.4%, respectively. The robustness of the Bayesian framework was assessed by comparing the inference performance of respective network models with that of the standard non-Bayesian equivalent, where networks are trained iteratively under variations injected into the weights following the canonical weight variation model [11]. Fig. 5(a), (b), (d), and (e) demonstrates the comparative inference results as bar chart representations for Bayesian and non-Bayesian counterparts under variations corresponding to different device sizes at VReact of 0.6 and 1.2 V, respectively. The robustness was evaluated under larger variations observed at a lower VReact of 0.6 V, where non-Bayesian trained networks suffer from substantial accuracy loss, which becomes more severe as the deviation increases with device downscaling. The more considerable accuracy degradation in AlexNet is primarily due to its deeper and more complex network architecture where variation across weights in all the layers gets accumulated to cause more ambiguity in the inference output. The proposed Bayesian framework dramatically minimizes the accuracy loss by retaining the near-ideal baseline accuracies for two shallow networks (MLP5 and LeNet) and exhibiting minimal accuracy drop for AlexNet. The accuracy Joss for AlexNet architecture on CIFARlO dataset has been observed to be 6.1%, 3.8%, and 16.I% with respect to the ideal accuracy value for 1/1 *µ,m,* 0.5/0.24 *µ,m,* and 0.24/0.24 *µm* sized devices, respectively, at VReact of 0.6 V. To make our study more meaningful, inference comparison results are also provided for a VReact of 1.2 V, as this voltage point is found to be optimal with respect to accuracy, training convergence, energy consumption, and conductance bit precision (as mentioned in the subsequent discussion). As expected, the smaller conductance variations for all device sizes at higher read voltages yield more improved accuracies. The result underscores the usefulness of our framework to provide robust and efficient inference under variations imposed even by highly scaled devices at smaller read voltages.

The benefits of adopting device-specific entire variation spectra as prior for the BNN (including interplay with device size and read voltage) instead of employing a uniform and fixed variation model for all network weights [12], [13] is substantiated by the accuracy comparison results provided in Fig. 5(a) and (c) and (d) and (f). The single variation value, *O'F,* used for each device size, as mentioned in

[1]Our implementation is based on a modified version of an open-source codebase available at httpsJ/github.com/kumar-shridhar/PyTorch-BayesianCNN.

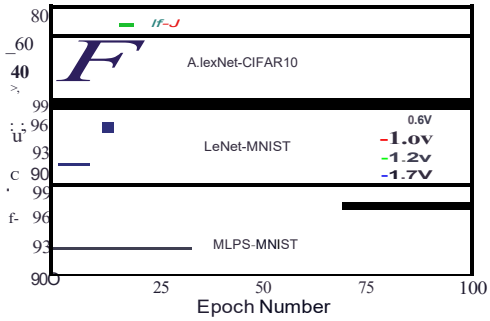

Fig. 6. Training dynamics of the proposed approach under variations corresponding to 0.24/0.24 μm device at different read voltages.

Fig. S(c) and (t), was estimated by averaging conductance variations over the entire operating range of the device [see Fig. 4(b)]. Our proposed scheme has been found to outperform the uniform variation-based method in terms of accuracy for all three network architectures exposed to different degree of weight fluctuations corresponding to different device sizes. It offers a noteworthy improvement in accuracy by 46.6%, 54.1%, and 35.2% for AlexNet-CIFARlO, LeNet-MNIST, and MLPS-MNIST, respectively, for the smallest device size operating at the lowest VRead-thereby substantiating the need for such hardware-software co-design efforts from a scalability perspective on complex machine learning tasks.

The impact of read voltage on the training dynamics of the proposed approach was also investigated for variations corresponding to the smallest device size. As revealed in Fig. 6, the higher read voltage not only improves the accuracy by offering lower conductance fluctuations but also accomplishes a stable training convergence at a relatively smaller number of epochs. However, the higher read operation causes more power consumption and limits the available number of conductance states in the operating range. Thus, a VRead of around 1.2 V could be an optimal solution to provide a reasonable trade-off between accuracy and power consumption.

V. CONCLUSION

In summary, we propose a novel device-algorithm codesign approach for reliable FE in-memory computing where a comprehensive conductance variation model derived by systematically characterizing FeFET devices is coupled to Bayesian learning-based uncertainty optimizer to alleviate the impact of device-level nonidealities. Incorporating dependencies of variation properties with operating voltage conditions and device size during the training process is shown to play a significant role in minimizing accuracy loss for complex datasets and deeper networks. The main advantage of this approach against hardware-in-the-loop training is that this will be a one-time training process without costly iterative training. Other process-variation-related issues like spatially correlated noise effects among neighboring FeFET devices in the crossbar array can be considered in future work to extend the efficacy of the proposed framework.

ACKNOWLEDGMENT

The authors would like to acknowledge GlobalFoundries, Dresden, Germany, for providing FeFET testing devices. They are thankful to Suma George Cardwell and Cale Douglas Crowder from Sandia National Laboratories for their valuable suggestions regarding the work.

REFERENCES

[I] I. Chakraborty, A. Jaiswal, A. **K.** Saha, S. **K.** Gupta. and K. Roy, "Pathways to efficient neuromorphic computing with non-volatile memory technologies," *Appl. Phys. Rev.,* vol. 7, no. 2, Jun. 2020, Art. no. 021308, doi: I0.1063/1.5113536.

[2] **W.** Haensch et al., "Compute in-memory with non-volatile elements for neural networks: A review from a co-design perspective," *Adv. Mater.,* vol. 35, no. 37, Sep. 2023, Art. no. 2204944, doi: I0.1002/adma.202204944.

[3] M Lederer et al., "Ferroelectric field effect transistors as a synapse for neuromorphic application," *IEEE Trans. Electron Devices,* vol. 68, no. 5, pp. 2295-2300, May 2021, doi: JO.I109rrED.2021.3068716.

[4] S. Deet al.,"28 nm HK.MG-based current limited FeFET crossbar-array for inference application," *IEEE Trans. Electron Devices,* vol.69, no. 12, pp. 7194-7198, Dec. 2022, doi: 10.1109rrED.2022.3216973.

[5] Y. Liu and P. Su, "Variability analysis for ferroelectric FET nonvolatile memories considering random ferroelectric-dielectric phase distribution," *IEEE Electron Device Lett,* vol. 41, no. 3, pp. 369-372, Mar. 2020, doi: JO.I 109/LED.2020.2967423.

[6] H. Mulaosmanovic et al., "Switching kinetics in nanoscale hafnium oxide based ferroelectric field-effect transistors," *ACS Appl Mater. Interfaces,* vol. 9, no. 4, pp. 3792-3798, Feb. 2017, doi: I0. I02 I/acsami.6b13866.

[7] B. Liu, H. Li, Y. Chen, X. Li, Q. Wu, and T. Huang, "Vortex: Variation-aware training for memristor X-bar," in *Proc. 52nd ACM/EDACIIEEE Design Autom. Co11f (DAC),* Jun. 2015, pp. 1-6, doi: J0.1145/2744769.2744930.

[8] L. Cben et al., "Accelerator-friendly neural-network training: Learning variations and defects in RRAM crossbar," in *Proc. Design, Autom. Test Eur. Conf Exhib. (DATE),* **Mar.** 2017, pp. 19-24, doi: J0.23919/DATE.2017.7926952.

[9] S. Jin, S. Pei, and Y. Wang, "On improving fault tolerance of memristor crossbar based neural network designs by target sparsifying," in *Proc. Design, Auto/IL Test Eur. Co11f Exhib. (DATE),* Mar. 2020, pp. 91-96, doi: J0.23919/DATE48585.2020.9116187.

[10] Y. Long, X. She, and S. Mukhopadhyay, "Design of reliable **DNN** accelerator with un-reliable **ReRAM,"** in *Proc. Design, Autom. Test Eur. Conf Exhib. (DATE),* 2019, pp. 1769-1774, doi: J0.23919/DATE.2019.8715178.

[11] Y. Bi, Q. Xu. H. Geng. S. Chen, and Y. Kang, "Resist: Robust network training for rnemristive crossbar-based neuromorphic computing systems," *IEEE Trans. Circuits Syst. II, Exp. Briefs,* vol. 70, no. 6, pp. 2221-2225, Jan. 2023, doi: JO. I109/TCSII.2023.3236168.

[12] D. Gao et al., "Bayesian inference based robust computing on memristor crossbar," in *Proc. 58th ACM/IEEE Design Autorn. Conf (DAC),* Dec. 2021, pp. 121-126, doi: JO.I109/DAC18074.2021.9586160.

[13] D. Gao et al., "BRoCoM: A Bayesian framework for robust computing on memristor crossbar," *IEEE Trans. Comput.-Aided Design huegr. Circuits Syst,* vol. 42, no. 7, pp. 2136--2148, Oct 2022, doi: JO.I109/TCAD.2022.3215071.

[14] M. Trentzsch et al., "'A 28 nm HK.MG super low power embedded **NVM** technology based on ferroelectric FETs," in *IEDM Tech Dig.,* Dec. 2016, pp. 1-4, doi: 10.1109/IEDM.2016.7838397.

[15] C. Li et al., "A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing," in *IEDM Tech Dig.,* Dec. 2020, pp. 29.3.1-29.3.4, doi: JO.I109/IEDM13553.2020.9372119.

[16] Y. Zhu et al., "Statistical training for neuromorphic computing using memristor-based crossbars considering process variations and noise," in *Proc. Design, Auto/IL Test Eur. Conf Exhib. (DATE),* **Mar.** 2020, pp. 1590-1593, doi: 10.23919/DATE48585.2020.9116244.

[17] A. Graves, "Practical variational inference for neural networks," in *Proc. Adv. Neural Inf Process. Syst.,* vol. 24, 201I, pp. 1-9.

[18] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat Assoc.,* vol. 112, no. 518, pp. 859--877, 2017, doi: 10.1080/01621459.2017.1285773.

[19] C. Blundell, J. Comebise, **K.** Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. 32nd Int. Conj Mach Leam,* Jun. 2015, pp. 1613-1622.

[20] X. Guo et al., "Ferroelectric FET-based strong physical unclonable function: A low-power, high-reliable and reconfigurable solution for Internet-of-Things security," 2022, *arXiv.2208./4678.*

[21] S. Thomann, A. Merna, K. Ni, and H. Amrouch, "Reliable FeFET-based neuromolJJhic computing through joint modeling of cycle-to-cycle variability, device-to-device variability, and domain stochasticity," in *Proc. IEEE Int. Rel. Phys. Symp. (/RPS),* **Mar.** 2023, pp. 1-5, doi: I0.1109/IRPS48203.2023. I0I 17810.

[22] **K. Ni, W.** Chakraborty, **J.** Smith, B. Grisafe, and S. Datta, "Fundamental understanding and control of device-to-device variation in deeply scaled ferroelectric FETs," in *Proc. Symp. VLSI Technol.,* Jun. 2019, pp. T40---T41, doi: I0.23919/VLSIT.2019.8776497.

[23] S. Chatterjee, S. Thomann, **K. Ni, Y.** S. Chauhan, and H. Amrouch, "Comprehensive variability analysis in dual-port FeFET for reliable multi-levekell storage," *IEEE Trans. Electron Devices,* vol. 69, no. 9, pp. 5316--5323, Sep. 2022, doi: I0.1109/TED.2022.3192808.

[24] C. Garg et al., "Impact of random spatial fluctuation in non-uniform crystalline phases on the device variation of ferroelectric FET," */EEE Electron Device Lett.,* vol. 42, no. 8, pp. 1160---1163, Aug. 2021, doi: I0.1109/LED.2021.3087335.

[25] **K. Ni,** A. Gupta, 0. Prakash, S. Thomann, X. S. Hu. and **H.** Amrouch, "Impact of extrinsic variation sources on the device-to-device variation in ferroelectric FET," in *Proc. IEEE Int. ReL Phys. Symp. (/RPS),* Apr. 2020, pp. 1-5, doi: I0.1109/IRPS45951.2020.9128323.

[26] C. Alessandri, P. Pandey, A. Abusleme, and A. Seabaugh, "Monte Carlo simulation of switching dynamics in polycrystalline ferroelectric capacitors," *IEEE Trans. Electron Devices,* vol. 66, no. 8, pp. 3527-3534, Jun. 2019, doi: I0.1109/fED.2019.2922268.

[27] S. Deng et al., "A comprehensive model for ferroelectric FET capturing the key behaviors: Scalability, variation, stochasticity, and accumulation," in *Proc. IEEE Symp. VLSI Teclmol.,* Jun. 2020, pp. 1-2, doi: I0.1109/VLSITechnologyl8217.2020.9265014.

[28] A. Saha, A. **N. M. N.** Islam, Z. Zhao, S. Deng, **K. Ni,** and A. Sengupta, "Intrinsic synaptic plasticity of ferroelectric field effect transistors for online learning," *Appl Phys. Lett.,* vol. 119, no. 13, Sep. 2021, Art no. 133701, doi: 10.1063/5.0064860.

[29] **A. N. M. N.** Islam, A. Saha, Z. Jiang, **K. Ni,** and A. Sengupta, "Hybrid stochastic synapses enabled by scaled ferroelectric field-effect transistors," *Appl. Phys. Lett.,* vol. 122, no. 12, Mar. 2023, Art. no. 123701, doi: I0.1063/5.0132242.

[30] Q. Xu, J. Wang, H. Geng, S. Chen, and X. Wen, "Reliability-driven neuromol}Jhic computing systems design," in *Proc. Design, Autom. Test Eur. Conj Exhib. (DATE),* Feb. 2021, pp. 1586--1591, doi: I0.239I9/DATE51398.2021.9473929.

[31] Y. LeCun, L. Bottou, Y. Bengio, and **P.** Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE,* vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: I0.1109/5.726791.

[32] A. Krizhevsky and G. Hinton. (2009). *Learning Multiple Layers of Features From Tiny Images.* [Online]. Available: https://www.cs.toronto.edu/-krizileaming-features-2009-TR.pdf

[33] A. Blum, N. Haghtalab, and A. D. Procaccia, "Variational dropout and the local repararneterization trick," in *Proc. Adv. Neural Inf Process. Syst. (NIPS),* 2015, pp. 2575-2583.

[34] S. Kullback and **R.** A. Leibler, "On information and sufficiency," *Ann Math. Statist,* vol. 22, no. I, pp.79--86, 1951. [Online]. Available: https://www.jstor.orwstablen236703