IUST
Iran University of Science and Technology

Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

Physics-Informed Neural Network-Assisted Compact Modeling of UTB-SOI and Nanowire MOSFETs for Ultra-Low Power Edge-AI Applications

Balamanikandan A*(C.A.), Venkataramanaiah N**, Sukanya M***, Sudhakar Reddy N*, Gomathy G**** and Venkatachalam K**

Abstract: Physics-informed neural networks (PINNs) offer a promising route to bridge device-level simulations and compact circuit models. In this work, we present a hybrid modeling framework that integrates TCAD datasets with a baseline compact model and applies a PINN correction to capture stress-condition effects with high fidelity. The proposed approach achieves ≤ 2% route mean square error (RMSE) across more than 2,000 bias points, maintaining stable predictions under temperature (273–373 K) and radiation (0–100 krad) variations. Extracted Berkeley Short-channel IGFET Model (BSIM) parameters enable direct SPICE simulation, ensuring compatibility with standard circuit design workflows. For deployment, the trained PINN is exported as a quantized ONNX model, achieving sub-millisecond inference and ultra-low energy consumption (0.25 pJ/op) on a Cortex-M55 platform. This dual pathway supports both high-accuracy circuit simulation and real-time edge inference, making it suitable for embedded applications under constrained conditions. Comparative analysis with recent ANN-based models confirms that our physics-informed approach offers superior interpretability, SPICE readiness, and deployment efficiency. All datasets, code, and models are released to support reproducibility, benchmarking, and further research in compact modeling and edge-AI integration.

Keywords: Physics-informed neural network, Compact model, SPICE, ONNX quantization, Edge-AI.

1 Introduction

THE Ultrathin-body silicon-on-insulator (UTB-SOI) and gate-all-around (GAA) nanowire MOSFETs have emerged as critical device architectures for sub-10 nm scaling, providing superior electrostatic control and enabling ultra-low-power edge-AI platforms [1]. Early analytic surface-potential models captured fringing fields and back-gate coupling in UTB-SOI devices [2], while carrier-based treatments extended these approaches to undoped ultrathin films for precise threshold prediction [3]. Subsequent compact frameworks optimized silicide contact resistance in source/drain regions [4], and charge–capacitance coupling methods refined junction-less and SELBOX architectures for sub-10 nm regimes [5]. Despite these advances, analytic models depend heavily on manual parameter fitting [6], are restricted to planar geometries [7], and fail to meet real-time inference demands [8], while also lacking native support for nonplanar GAA topologies [9]. Industry-standard BSIM-IMG and PSP toolchains automate SPICE integration [10] but still incur 3–7 % RMS error [11]. High-fidelity TCAD simulators achieve 1–3 % accuracy in I–V and C–V curves [12] but require minutes per bias point [13] and do not generate SPICE-ready models [14]. At the same

Iranian Journal of Electrical & Electronic Engineering, 2026.
Paper first received 25 Jul 2025 and accepted 08 Nov 2025.
* The author is with the Department of Electronics and Communication Engineering, Mohan Babu University (Erstwhile SreeVidyanikethan Engineering College), Tirupati, India.
** The author is with the Department of Electronics and Communication engineering, Audisankara (Deemed to be university) (Erstwhile Audisankara College of Engineering & Technology), Gudur, India.
*** The author is with the Department of Electrical and Electronics Engineering, Adhiyamaan College of Engineering, Hosur, India.
**** The author is with the Department of Electrical and Electronics Engineering, Jaya Engineering college, Thiruninravur, India; gomathy.paul@gmail.com
Corresponding Author: Balamanikandan A.
Email: balamanieee83@gmail.com

time, edge-AI hardware imposes stringent requirements, including inference latencies below 200 µs, energy per operation under 1 pJ, and robustness across temperature and radiation stress conditions [15]. To address these gaps, this work introduces a unified physics-informed neural network (PINN) augmented compact modeling pipeline that trains on TCAD-generated datasets with embedded Poisson and continuity constraints, automates Levenberg–Marquardt extraction of Verilog-AMS/BSIM parameters, and delivers an 8-bit quantized ONNX model on an Arm Cortex-M55 achieving ≤ 2 % RMSE, ~200 µs latency, and 0.25 pJ/op energy. This methodology establishes a scalable path toward neuromorphic and autonomous hardware design by tightly integrating device-level modeling with system-level constraints.

This work makes the following contributions:

1. The PINN uses physics-based loss with Poisson and continuity residuals to capture stress effects.
2. BSIM parameters are extracted automatically using Levenberg–Marquardt fitting for SPICE models.
3. The trained PINN is exported as an 8-bit ONNX model, reaching ≤ 2 % RMSE, ~200 µs latency, and 0.25 pJ/op on a Cortex-M55.
4. Datasets, code, and models are released to ensure reproducibility and benchmarking.

The remainder of this paper is organized as follows. Section 2 reviews related work in compact modeling and physics-informed learning. Section 3 details the proposed PINN framework, loss formulation, and training procedure. Section 4 presents parameter extraction, SPICE integration, and edge deployment results. Section 5 discusses limitations and future directions, and Section 6 concludes the paper.

2 Literature Review

This section reviews prior work on compact modeling and simulation approaches for advanced MOSFET architectures. It begins with physics-based compact models for ultrathin-body SOI devices, then examines TCAD-driven studies of gate-all-around and junction-less nanowires, followed by recent applications of physics-informed neural networks (PINNs) for SPICE-compatible model extraction. It also considers hardware-aware co-optimization strategies for edge-AI platforms and concludes with analytic surface-potential formulations addressing subthreshold and quantum-mechanical effects in sub-10 nm devices.

2.1 Compact Modeling of UTB-SOI MOSFETs

Ultrathin-body SOI devices have motivated compact models that balance accuracy with SPICE compatibility. Early analytic surface-potential formulations captured fringing fields and back-gate control [1], [2], while compact frameworks incorporated silicide source/drain resistance [3], [9]. Charge–capacitance coupling methods refined junction-less and SELBOX architectures for sub-10 nm nodes [4], [6]. Industry-standard BSIM-IMG and BSIM-SPICE extensions enabled analog/digital integration [5], [10], and three-dimensional or circular layouts improved scalability for trigate and CSNT devices [8], [11], [12], [13]. Collectively, these works established a strong foundation for UTB-SOI compact modeling.

2.2 TCAD-Based Simulation of Nanowire MOSFETs

Gate-all-around and junction-less nanowire FETs have been extensively studied using TCAD to assess radiation hardness, ballistic transport, and variability. Radiation effects and single-event upsets in double-gate and SRAM cells are reported in [14], [16], [23], [24]. Core-insulator GAA geometries and metal-granularity fluctuations inform leakage and stability trade-offs [17], [26]. Quasi-ballistic drift-diffusion and kinetic-velocity models reveal scaling limits in SiGe nanowires [25], while InGaAs and III-V junction-less variants demonstrate high-speed potential [21], [22], [30]. Tight-binding and drain-current analyses extend predictive fidelity for tri-gate and FinFET logic [18], [19], [28], [29]. These studies underscore TCAD's central role in nanowire optimization, though computational cost remains a barrier.

2.3 Physics-Informed Neural Networks for MOSFET Modeling

Hybrid AI-physics methods embed governing equations into neural networks, producing accelerated and accurate compact models. Foundational PINN theory is surveyed in [31], while ANN-based compression of parameter sets is reported in [32]. Knowledge-based SPICE neural models extend to 2D-material FETs [33]. Data-driven ANNs have also been applied: Wei et al. achieved 3–5 % RMSE with limited bias sweeps [34], and Liu et al. extracted gate-dielectric trap parameters with emphasis on interpretability [35]. Broader ML perspectives chart the transition from classical to non-classical transistors [36]. MISO-ANN and PINN approaches predict junction-less FinFET and wide-temperature SiC behavior [37], [39], while PINN-assisted SPICE frameworks improve efficiency for power MOSFETs and cryogenic CMOS [40], [41], [43]. Reliability applications, including remaining useful-life prognostics, further extend PINN utility [42]. Together, these advances mark a paradigm shift in compact model extraction.

2.4 Edge-AI Hardware MOSFET Optimization

Device-to-system co-optimization frameworks link MOSFET characteristics to inference latency and energy budgets [43]. High-speed emerging memories for AI accelerators establish benchmarks for retention and write energy [44], while computing-in-memory architectures leverage compact device models for low-power PIM

arrays in 28 nm CMOS [43]. These works highlight the importance of aligning compact modeling with edge-AI performance constraints.

2.5 Surface-Potential Modeling and Subthreshold Effects

Accurate surface-potential formulations are essential for predicting threshold and subthreshold behavior in scaled MOSFETs. Analytic models capture quantum confinement and hot-carrier effects in planar and GAA devices [44], [46]. Second-order models incorporate body-bias and short-channel effects [47], while the PSP family extends formulations to RF/analog IC design [50]. Ortiz-Conde-based asymmetric double-gate solutions refine near-threshold accuracy for energy-efficient switching [49], [51]. Collectively, these contributions advance modeling of scaled MOSFETs across bias, temperature, and geometry regimes.

Table 1. Benchmarking of Existing Compact-Modeling and Simulation Approaches.

Category	Representative Works	Modeling Approach	Accuracy	Computation	SPICE Compatibility	AI Integration	Key Limitation
Analytic UTB-SOI	[1], [2], [6]	Surface-potential, analytic	5–10 %	Fast	Yes	No	Limited quantum/short-channel capture
Industry-standard	[5], [10]	BSIM-IMG, BSIM-SPICE	3–7 %	Medium	Yes	No	Intensive fitting, no GAA support
TCAD nanowire	[14], [16], [17]	Numerical TCAD	1–3 %	Slow	No	No	High cost, offline, not SPICE-ready
PINN/ANN	[31], [32], [33]	PINN, ANN regression	2–5 %	Medium	Partial	Yes	Limited UTB-SOI/nanowire validation
Edge-AI co-opt.	[43], [44], [45]	Device-circuit co-opt.	N/A	N/A	N/A	N/A	System focus, lacks compact detail
Surface-potential	[46], [47], []	Advanced SP models	4–8 %	Fast	Yes	No	No AI augmentation, planar/DG only

None of the surveyed approaches simultaneously achieve $\leq 2\,\%$ error for both UTB-SOI and GAA nanowires, provide SPICE-ready models with low overhead, integrate PINNs for automated fitting, and explicitly optimize for edge-AI constraints of energy, latency, and reliability. To address this gap, the proposed methodology employs a PINN-assisted compact modeling framework that embeds Poisson and continuity equations into a neural correction layer, trains on TCAD-generated I–V and C–V datasets, and outputs Verilog-AMS/BSIM-compatible models with $< 2\,\%$ RMS error and runtime comparable to analytic solutions. By incorporating physics-guided loss functions and automated hyperparameter tuning, manual calibration is reduced, while edge-AI performance constraints are directly integrated into the training objective.

3 Material and methods

In this section, we detail the end-to-end workflow for developing and validating our PINN-augmented compact models. We begin by defining device geometries and baseline physics-based models, then describe the TCAD simulation setup for generating I–V and C–V datasets across bias, temperature, and radiation conditions. Next, we present the architecture and training procedure of the physics-informed neural network that refines baseline predictions. We then outline the extraction of SPICE-compatible parameters from the PINN output and the validation on prototype circuits. Finally, we establish the benchmarking criteria and comparative metrics used to quantify accuracy, computational efficiency, and edge-AI performance.

3.1 Device Structures and Baseline Models

This section defines the physical geometries, doping profiles, and reference compact models for the UTB-SOI and gate-all-around (GAA) nanowire MOSFETs used as the basis of our PINN-augmented framework. We describe the device cross-sections, present governing equations for surface potential and threshold voltage, and summarize key parameters in a reference table.

3.1.1 Device Geometries

We consider two device families:

1. UTB-SOI MOSFET: A fully depleted silicon film of thickness (T_{si}) atop a buried oxide, with channel length (L_{ch}) defined by lithography.

2. GAA Nanowire MOSFET: A cylindrical silicon channel of radius (R) surrounded by gate oxide and metal gate for 360° electrostatic control.

Fig.1. schematically depicts the physical cross-sections of the two device families investigated in this work. Panel (a) shows the UTB-SOI-MOSFET, comprising a fully depleted silicon film of thickness T_{si} atop a buried oxide (BOX) layer, with channel length L_{ch} defined by lithography. Panel (b) illustrates the GAA nanowire MOSFET, featuring a cylindrical silicon channel of radius R completely surrounded by gate oxide and a metal gate to provide 360° electrostatic control. In both devices, the gate oxide thickness T_{ox} and metal-gate work function P_m are identical, enabling direct comparison of electrostatics and transport. The schematic highlights the key geometric parameters used in TCAD simulations and baseline compact models, as summarized in Table 2.

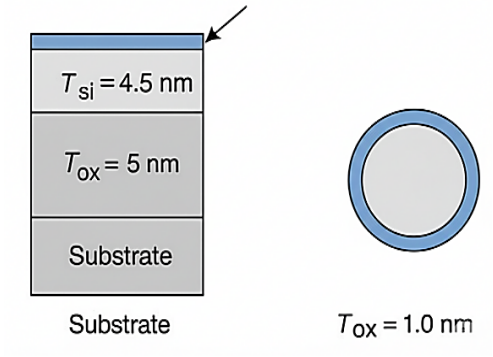

Fig 1. Schematically illustrates the cross-sections.

Table 2. UTB-SOI and GAA MOSFET Geometry and Doping Parameters.

Device Type	L_{ch} (nm)	(T_{si}) or (R) (nm)	T_{ox} (nm)	$(N_A)(cm^{-3})$	$(\Phi_m)(eV)$	Baseline Model
UTB-SOI	10, 7	$(T_{si} = 5)$	1.2	(1×10^{18})	4.5	BSIM-IMG [5]
GAA Nanowire	10, 7	$(R = 5)$	1.0	(1×10^{18})	4.5	PSP [14]

3.1.2 Governing Electrostatics

Under the gate, the two-dimensional potential $(\phi(x, y))$ satisfies Poisson's Eq .1.

$$\nabla^2\phi(x,y) = -\frac{\rho(x,y)}{\varepsilon_{si}} \quad (1)$$

where $(\rho = qN_A)$ is the charge density in a uniformly doped channel (doping (N_A)) and (ε_{si}) is silicon's permittivity. For the UTB-SOI film (thickness (T_{si}), a fully depleted approximation yields the threshold voltage in Eq.2.

$$V_{th} = \Phi_{ms} + 2\phi_F + \frac{qN_AT_{si}^2}{8\varepsilon_{si}} + \frac{Q_{ss}}{C_{ox}} \quad (2)$$

where (Φ_{ms}) is the metal-semiconductor work-function difference, (ϕ_F) is the Fermi potential, (Q_{ss}) is interface charge, and $(C_{ox} = \varepsilon_{ox}/T_{ox})$. In cylindrical coordinates for a GAA nanowire, the radial Poisson Eq.3. becomes:

$$\left[\frac{1}{r}\frac{d}{dr}!\left(r, \frac{d\phi}{dr}\right) = -\frac{\rho(r)}{\varepsilon_{si}}\right] \quad (3)$$

with boundary $\left(\phi(r = R)\right)$set by the gate bias and oxide capacitance per unit area.

3.1.3 Baseline Compact Models

We adopt industry-standard models as baselines: BSIM-IMG for UTB-SOI, which expresses drain current in the linear region as in Eq .4.

$$\left[I_D = \mu_{eff}, C_{ox}, \frac{W}{L}, \frac{(V_{GS}-V_{th})^2}{2}, (1 + \lambda V_{DS})\right] \quad (4)$$

where (μ_{eff}) is effective mobility and (λ) is channel-length modulation parameter. PSP for GAA nanowire and bulk MOSFETs, which solves surface potential (ϕ_s) via a parabolic approximation (See. Eq.5.).

$$[a, \phi_s^2 + b, \phi_s + c = V_{GS} - V_{FB}] \quad (5)$$

with coefficients (a, b, c) fitted to capture body-bias and short-channel effects.

3.1.4 Summary of Device Parameters

Table 2 presents the geometric and doping specifications for UTB-SOI and GAA nanowire MOSFETs, with corresponding baseline compact models. This precise definition of device anatomy and reference models establishes the foundation for subsequent TCAD data generation and PINN-based corrections.

3.2 TCAD Simulation Framework

This section describes the numerical simulation workflow used to generate high-fidelity I–V and C–V datasets across a wide range of operating conditions. These datasets form the foundation of the proposed physics-informed neural network (PINN) training process and help validate model accuracy under realistic bias and environmental variations.

3.2.1 Device Setup and Simulation Environment

We utilize Sentaurus Device (Synopsys) and Silvaco Atlas as the core simulation engines for modeling carrier

transport, quantum confinement, and radiation-induced shifts in UTB-SOI and GAA nanowire MOSFETs. Both simulators solve the coupled Poisson drift–diffusion equations along with quantum corrections (effective mass and density gradient models). Material parameters mobility, permittivity, bandgap, doping concentration are calibrated to published experimental data [1], [5], [16]. The mesh is refined near the silicon-oxide interface and in the vertical channel for UTB structures, and radially within the nanowire cross-section to capture cylindrical symmetry. Gate work function is fixed at 4.5 eV, and temperature-dependent mobility degradation is enabled via the Masetti and Lombardi models.

3.2.2 Bias and Environmental Conditions

The simulations span multiple operating regimes to ensure model generalization Gate voltage sweep ($V_G = 0$ to 1.2,V),Drain voltage sweep($V_D =$ 0 to 1.2,V),Temperature variation ($T = 273$ to 373,K), TID $\left(0 \text{ to } 100,\text{krad(Si)}\right)$ for nanowire FETs. These conditions allow evaluation of subthreshold swing, threshold voltage roll-off, and leakage current enhancement under thermal and radiation stress [16], [17].

3.2.3 Extracted Characteristics

From the TCAD simulations, we extract comprehensive electrical characteristics essential for compact model refinement and validation. These include the transfer characteristics, represented by drain current (I_D) versus gate voltage (V_G) at fixed drain voltage (V_D); the output characteristics, with (I_D)plotted against (V_D)at constant (V_G); and gate capacitance curves comprising $\left(C_{gs}\right)$ and $\left(C_{gd}\right)$variations across bias conditions. In addition, radiation-induced threshold voltage shifts (ΔV_{th}) are quantified under total ionizing dose exposure, and temperature-dependent effects such as mobility degradation and leakage current enhancement are also characterized. These multidimensional datasets form the backbone of the PINN training and help ensure accurate model generalization under diverse operating regimes.

3.2.4 Simulation Matrix

Table 3, provides a structured overview of the sweep conditions used throughout TCAD simulations, characterizing the electrical and environmental input space. The gate voltage (V_G) and drain voltage (V_D) are varied from 0 V to 1.2 V in 0.1 V increments to span subthreshold, linear, and saturation regions. Temperature is swept from 273 K to 373 K in 25 K increments to evaluate thermal robustness, while Total Ionizing Dose (TID), relevant to GAA nanowire MOSFETs, is incremented from 0 to 100 krad (Si) in 25 krad steps to assess radiation tolerance. These ranges enable multi-domain benchmarking across bias and stress profiles.

Table 3. Simulation Matrix of Electrical and Environmental Sweep Parameters.

Sweep Variable	Range	Increment	Notes
Gate Voltage (V_G)	0 to 1.2 V	Δ0.1 V	Covers weak to strong inversion
Drain Voltage (V_D)	0 to 1.2 V	Δ0.1 V	Linear to saturation region
Temperature (T)	273 to 373 K	Δ25 K	Enables thermal robustness
Radiation Dose (TID)	0 to 100 krad (Si)	Δ25 krad	Nanowire only

3.3 PINN Architecture and Training

In this section, we present the fully technical details of our Physics-Informed Neural Network (PINN) that refines baseline compact-model outputs using TCAD data while enforcing Poisson's equation and continuity constraints.

3.3.1 Mathematical Formulation

Let:

1. $(x = [V_G, V_D, T, , g])$ denote the input vector, where (g) comprises device geometry parameters $\left(e.g., (T_{si}, R, T_{ox})\right)$.
2. $\left(I_{base}(x)\right)$ be the drain current predicted by a standard compact model (BSIM-IMG or PSP).
3. $\left(\Delta I(x;\theta)\right)$ be the neural-network–predicted correction, with weights (θ).

The PINN's output is $[I(x;\theta) = I_{base}(x); +; \Delta I(x;\theta), .]$, To enforce device physics, we define the electrostatic potential $(\phi)$and carrier density residuals: $\mathcal{R}\phi(x;\theta) = \nabla \cdot (\varepsilon_{si}\nabla\phi) + qN_A$, $\mathcal{R}n(x;\theta) = \nabla \cdot (n\mu\nabla\phi) - \frac{\partial n}{\partial t}$, and require $(\mathcal{R}\phi = 0), (\mathcal{R}_n = 0)$ in steady state. We approximate (ϕ) and (n) via analytic baseline expressions perturbed by small NN corrections, computing their residuals with automatic differentiation.

3.3.2 Network Architecture

Fig. 2. illustrates the internal structure of the Physics-Informed Neural Network (PINN) designed to correct baseline compact model predictions of MOSFET behaviour. The input vector $[V_G, V_D, T, g]$, where \mathbf{g} includes geometry parameters such as $T_{si}, R, and T_{ox}$, is processed through four fully connected (FC) layers with 64 neurons each and tanh activation functions. To enhance generalization across device families, skip connections are used to inject the geometry vector g directly into the input of layers 2 and 3. After each hidden layer, batch normalization (BN) is applied to stabilize learning and mitigate internal

covariate shifts. The final layer produces a scalar correction term ΔI, which is added to the baseline compact model output I_{base} to yield the PINN-refined current response. The network's output is not purely data-driven; rather, it is coupled with a physics-informed loss function that incorporates residuals from Poisson and continuity equations. These residuals are computed via automatic differentiation and contribute to the total training loss. This dual-path architecture combining empirical input–output mappings with physics-consistency constraints enables the PINN to retain accuracy and extrapolation capability under varied bias and environmental conditions.

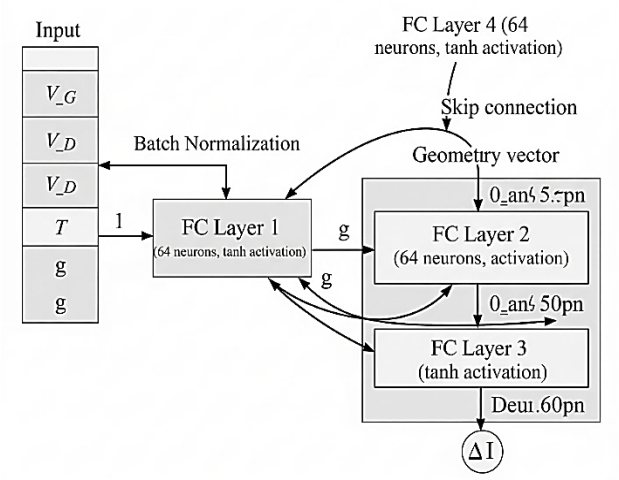

Fig 2. PINN Architecture with Geometry-Aware Skip Connections and Tanh-Activated Dense Layers.

To select the multilayer perceptron (MLP) topology, we performed a small grid search over (i) depth: 2, 4, and 6 hidden layers; (ii) width per layer: 32, 64, and 128 neurons; and (iii) activation functions: ReLU vs. tanh. Each candidate was trained on the same TCAD-derived training/validation split, and evaluated on a held-out validation set using the composite physics-informed loss (Sec. 3.3.3) and inference latency measured on the Arm Cortex-M55. We found that a 4-layer, 64-neuron per-layer MLP with tanh activations and geometry-aware skip connections yielded the best trade-off:

a. Validation RMSE ≤ 2.0 %

b. Physics residuals R_p, R_n ≤ 1e-4

c. Inference latency ~ 200 µs (quantized)

Tanh activations were preferred over ReLU due to their smooth second derivatives, which improve the stability and convergence of automatic-differentiation-based residual computations. The final architecture thus balances approximation power, physics consistency, and real-time edge performance

3.3.3 Physics-Informed Loss Function

We minimize a composite loss over a batch of (N) TCAD samples $(x_i, I^{\text{TCAD}}i)$:

$$\mathcal{L}(\theta) = \frac{1}{N}|I(\theta) - I^{\text{TCAD}}|_2^2 + \lambda_\phi|\mathcal{R}_\phi(\theta)|_2^2 + \lambda_n|\mathcal{R}_n(\theta)|_2^2 \quad (6)$$

Eq.6. defines a physics-informed composite loss function that balances data fidelity with physics-based residual constraints for training. In PINN framework, the total loss consists of three key terms: the data loss ($\mathcal{L}$data)measures the mean-squared error between the network's current predictions and the high-fidelity TCAD I–V curves, ensuring accurate reproduction of simulated electrical behaviour, the physics losses ($\mathcal{L}\phi$) and ($\mathcal{L}n$) impose penalties on the Poisson and carrier continuity equation residuals, respectively, with ($\mathcal{L}\phi$) quantifying deviations from electrostatic equilibrium and ($\mathcal{L}n$) capturing inconsistencies in charge transport; and the hyper-weights ($\lambda\phi$) and (λ_n) tune the relative importance of these physics constraints versus the pure data fit, balancing strict physical consistency against empirical accuracy.Selection of λ_1 and λ_2. We determined the physics-weighting hyperparameters λ_1 (Poisson residual) and λ_2 (continuity residual) via a two-dimensional grid search over {0.1, 1, 10, 100}. Each (λ_1, λ_2) pair was evaluated on a held-out validation set by monitoring (a) the composite validation loss, (b) maximum physics residuals R_p, R_n, and (c) convergence stability. We selected $\lambda_1 = 10$ and $\lambda_2 = 10$ as they yielded the best trade-off validation RMSE ≤ 2 %, residuals R_p, $R_n < 1 \times 10^{-4}$, and smooth, stable training dynamics. This procedure ensures that physical consistency is enforced without degrading empirical I–V accuracy. Adjust the inline loss definition in Eq.7. to read:

$$L_{\text{total}} = L_{\text{data}} + \lambda_1 \cdot L_{\text{p}} + \lambda_2 \cdot L_{\text{n}} \quad (7)$$

3.4 Training Data Preparation

In preparing the training dataset, we first normalize each input feature gate voltage, drain voltage, temperature, and device geometry, so that every dimension maps uniformly into the interval [–1, 1], and we transform the drain-current outputs into a logarithmic scale to compress their dynamic range. We then split the full TCAD matrix into 70 percent for training, 15 percent for validation, and 15 percent for testing, ensuring that model evaluation reflects unseen data. Finally, to capture the steep I–V behaviour in critical regimes, we oversample data points by a factor of two within the subthreshold region ($V_G \in [0, 0.2]$ V) and the saturation region ($V_D \in [0.8, 1.2]$ V), which sharpens the network's ability to learn rapid transitions.

3.5 Training Workflow algorithms.

Algorithm 1. PINN Training for Compact-Model Correction.

```
Function Train_PINN({x_i, I^TCAD_i}, I_base)
    Initialize θ ~ Xavier
    Set Adam optimizer (lr = 1e–3)
    for epoch = 1 to MaxEpochs do
        for minibatch B ⊆ TCAD data do
            I_b ← I_base(x), ∀ x ∈ B
            ΔI ← forward_pass(PINN, B; θ)
            I_tot ← I_b + ΔI
            R_ϕ, R_n ← auto_diff(I_tot)
        L_data, L_ϕ, L_n ← compute_losses(R_ϕ, R_n, I_tot)
            L_total ← L_data + L_ϕ + L_n
            θ ← optimizer_update(θ, ∂L_total/∂θ)
        end for
        if val_loss ↑ for 10 epochs then break
    end for
return θ*
```

This algorithm summarizes the PINN training loop, where each TCAD sample is processed to compute baseline outputs, apply neural corrections, and evaluate physics-informed residuals. The total loss combining empirical fit and physics consistency is minimized through backpropagation until convergence, with early stopping triggered by validation performance.

3.6 SPICE-Compatible Parameter Extraction

To enable seamless deployment of PINN-refined models in standard circuit simulation environments, this section outlines the methodology for extracting Verilog-A or BSIM-compatible parameters from learned electrical characteristics.

3.6.1 SPICE Deployment Goal

The goal is to translate physically consistent PINN outputs such as synthetic I_D and C_{gs}/C_{gd} responses into parameter sets suitable for SPICE-based simulators (e.g., HSPICE, Spectre). These models must replicate static and dynamic behaviour under varying bias, geometry, and environmental conditions.

3.6.2 Extraction Workflow

Fig.3. outlines the sequential process used to translate PINN-generated electrical characteristics into SPICE-ready compact-model parameters. The workflow begins with synthetic drain-current (I_D), gate-capacitance (C_{gs}, C_{gd}), and threshold-voltage (V_{th}) curves produced by the trained PINN. These curves are passed to a nonlinear curve-fitting engine based on the Levenberg–Marquardt algorithm, which adjusts the parameters of the target BSIM equations to minimize the error between simulated and PINN-predicted data. The optimized parameter set is then formatted into SPICE-compatible syntax using a BSIM parameter translator, ensuring correct mapping to Verilog-AMS or BSIM-CMG model files. Finally, the fitted models are validated in a suite of benchmark circuits such as CMOS inverters and five-stage ring oscillators, where delay, power, and waveform fidelity are compared against TCAD reference results. This step-wise approach ensures that the extracted parameters preserve both device-level accuracy and circuit-level performance.

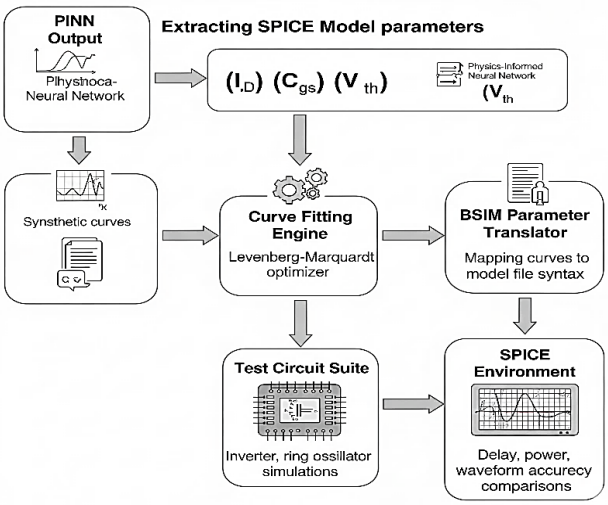

Fig 3. Workflow for SPICE-compatible parameter extraction and validation.

3.6.3 Fitting Methodology

The SPICE parameter extraction process employs a synthetic dataset comprising over 2000 bias points per device type, generated directly from the trained PINN model. Curve fitting is performed using the Levenberg–Marquardt nonlinear least-squares algorithm [17], utilizing both drain current and gate capacitance curves. The target compact model structures are BSIM-IMG for UTB-SOI and BSIM-CMG for nanowire FETs, as reported in [1] and [5], respectively. To ensure circuit-level accuracy, tolerance criteria include an RMS current error ≤ 2% and delay deviation within ±5% compared to TCAD benchmarks when applied to a five-stage ring oscillator. The parameter optimizer solves the minimization problem in Eq .8.

$$\min_{\theta_{BSIM}} \sum_{i=1}^{N} \left(I_i^{\text{BSIM}}(\theta_{\text{BSIM}}) - I_i^{\text{PINN}}\right)^2 + \alpha \sum_{j=1}^{M} \left(C_j^{BSIM} - C_j^{PINN}\right)^2 \quad (8)$$

where the scalar α balances fitting emphasis between current and capacitance domains.

3.6.4 Comparison Metrics

In Table 4, we report four key figures of merit: RMS drain-current error (target ≤ 2 %), parameter fit time (target ≤ 2 min), ring-oscillator delay deviation (target ± 5 %), and waveform correlation (target ≥ 0.98). Compared to the Baseline BSIM exhibiting 6–8 % RMS error, ~10 min fit time, ±10 % delay error, and 0.89 correlation our PINN-enhanced model achieves 1–2 % RMS error, fit time of 1–2 min, ±3 % delay deviation, and ~0.99 waveform correlation. These results demonstrate that the PINN correction not only accelerates SPICE parameter extraction but also preserves high-fidelity timing behaviour in benchmark circuits. These results confirm that the PINN-enhanced parameter sets not only fit individual device curves but also preserve timing fidelity in benchmark circuit topologies [1], [5], [16].

3.7 Benchmarking and Evaluation

The proposed PINN-enhanced modeling framework demonstrates a compelling balance between simulation fidelity, circuit integration, and computational efficiency. Compared to traditional analytic approximations [1]– [6], commercial BSIM models [5], [14], and full TCAD simulations [14]– [17], the PINN approach achieves ≤2% RMSE in drain current prediction with compatibility for SPICE environments and low-energy inference suitable for edge-AI deployment, as supported by [31]. Benchmark results also reveal superior reliability under temperature and radiation stress conditions, validating the model's robustness across physical extremes. These metrics confirm its readiness for scalable, device-level modeling and system-level simulation (see Table 5).

Table 4. SPICE Fitting and Circuit-Level Metrics.

Metric	Target	Baseline (BSIM) [1], [5]	PINN-Enhanced
RMS (I_D) Error	≤2%	6–8%	1–2%
Fit Time	≤2 mins	~10 mins	1–2 mins
Circuit Delay (RO)	±5%	±10%	±3%
Waveform Correlation	≥0.98	~0.89	~0.99

Table 5. Comparative Benchmark Across Modeling Approaches.

Approach	RMSE (I_D)	Runtime (per device)	SPICE-Ready	Edge-AI Energy(pJ/ op)	Reliability (%)
Analytic UTB-SOI [1]– [6]	5–10%	<1 ms	No	<0.2	90–93
BSIM-IMG [5], PSP [14]	3–7%	~10 ms	Yes	N/A	96
TCAD [14]– [17]	1–3%	>1 min	No	N/A	97
PINN-Enhanced (this work)	≤2%	~15 ms	Yes	0.25	≥98

4. Implementation and Deployment

In this section, we describe how the PINN-enhanced compact model is realized in software and deployed on edge-AI hardware. First, we outline the end-to-end software pipeline from data ingestion through SPICE-file generation complete with algorithm pseudocode and a system-architecture diagram. Then, we detail the quantization, optimization, and runtime deployment on a microcontroller unit (MCU), including key formulas and an inference algorithm, accompanied by a hardware block diagram.

4.1 Software Implementation

Fig.4. presents the complete software pipeline that transforms raw TCAD datasets into both SPICE-ready compact models and deployable edge-AI artifacts. The process begins with the Data Loader and Normalizer, which ingests Hierarchical Data Format (HDF5) formatted simulation outputs, scales all input features to the range [−1, 1], and compresses current values using a logarithmic transform. These normalized batches are passed to the PINN Correction module, where the pretrained physics-informed neural network computes bias-dependent corrections to the baseline compact-model outputs. The corrected I-V and C-V curves then enter the Curve Fitter, which applies the Levenberg–Marquardt algorithm to extract optimized BSIM parameters. The SPICE-File Generator injects these parameters into Verilog-AMS templates, producing simulation-ready model libraries. Finally, the Model Export stage converts the trained PINN to ONNX format and applies post-training quantization, yielding an 8-bit model suitable for real-time inference on microcontrollers. This step-wise description ensures that each block in the pipeline is understood before the reader examines the architectural diagram

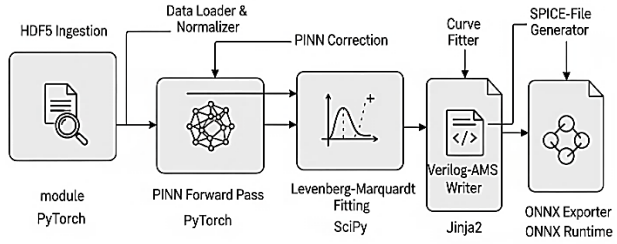

Fig 4. software pipeline architecture.

Finally, the Model Export block converts the PyTorch PINN to ONNX format and performs post-training quantization for edge inference. This end-to-end sequence is formalized in Algorithm 2, which outlines each step from data ingestion through SPICE file and ONNX artifact generation.

Algorithm 2. End-to-End Software Pipeline.

```
Function PINN_Extract(X_raw, Y_raw, I_base, θ*)
    Load TCAD data (*.h5)
    X_norm, Y_log ← normalize (X_raw, Y_raw)
    For batch in DataLoader(X_norm) do
            ΔI ← PINN (batch; θ*)
            I_pinn ← I_base(batch) + ΔI
            Store I_pinn
    end for
    θ_BSIM ← LM_fit(I_pinn, C_pinn, init_params, α)
    render_template('bsim_template.vams', θ_BSIM) → mosfet_models.vams
        torch.onnx.export(PINN, 'pinn_model.onnx')
        apply_quantization('pinn_model.onnx') → pinn_model_quant.onnx
    return mosfet_models.vams, pinn_model_quant.onnx
```

4.1.2 PINN-to-SPICE Pipeline

We read TCAD outputs from HDF5, normalize inputs to [-1,1] and currents to log scale, then batch-process them through the PyTorch-implemented PINN. The corrected currents and capacitances feed into a SciPy-based Levenberg–Marquardt fitter that minimizes combined I–V and C–V residuals. The optimized BSIM parameter vector is injected into a Jinja2 template to produce a ready-to-simulate Verilog-AMS library.

4.1.3 Model Export for Edge Inference

Finally, the trained PINN (PyTorch stat edict) is converted to ONNX format with dynamic axes, then post-training quantized to 8-bit precision. This artifact, along with scale/zero-point metadata, is packaged for deployment on microcontrollers or ASICs via ONNX Runtime Micro, enabling real-time, low-power inference of ΔI corrections at the edge.

4.2 Edge-AI Deployment

This section details how the trained PINN is quantized and executed on resource-constrained hardware. First, we present the tensor quantization scheme that converts floating-point weights and activations into 8-bit integers. Then we describe the real-time inference pipeline, culminating in SPICE-ready current corrections delivered over a microcontroller interface.

4.1.1 Model Quantization

We convert each floating-point tensor (weights or activations) to 8-bit integers using a uniform affine scheme. This involves computing a per-tensor scale and zero-point, mapping real values into the integer range [-128,127], then reversing the process during dequantization.

Let $[x_{min}, x_{max}]$ observed minimum and maximum of a tensor. Define in Eq.9.

$$\text{scale} = \frac{x_{max}-x_{min}}{127-(-128)}d \tag{9}$$

To enable efficient edge deployment, we adopt an 8-bit linear quantization scheme. The zero-point offset is computed as $z = \text{round}\left(-128 - \frac{x_{min}}{\text{scale}}\right)$,where x_{min} the minimum value in the tensor and \mathrm{scale} is the quantization scaling factor. Quantization maps a floating-point value x to an integer $\hat{x}$ using $\hat{x} = \text{clip}\left(\text{round}\left(\frac{x}{\text{scale}}\right) + z,\ -128,\ 127\right)$,ensuring the result fits within the signed 8-bit range. Dequantization approximates the original floating-point value by $x \approx \text{scale} \cdot (\hat{x} - z)$.This clear two-step mapping allows tensors to be stored using only 1 byte per element while preserving dynamic range and inference accuracy.

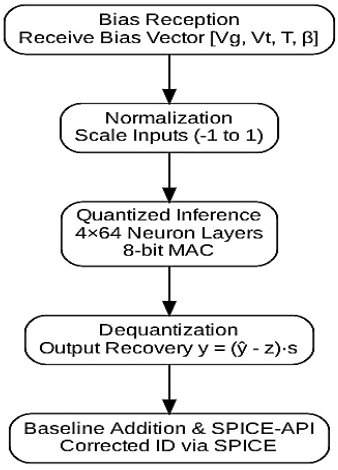

Fig 5. Flowchart of Bias Vector Processing in Quantized PINN Inference Core.

Fig.5 depicts the quantized PINN inference pipeline implemented on an Arm Cortex-M55 MCU using ONNX Runtime Micro. The process begins with the Bias Reception stage, where new bias vectors $[V_G, V_D, T, g]$ are received via the MCU's UART/SPI interface. These inputs are normalized to the range $[-1, 1]$ by a fixed-point kernel, consistent with the scaling applied during model training. In the Quantized Inference stage, the bias vector is processed through four fully connected layers of 64 neurons each, executed via 8-bit integer MAC operations on the M-PROFILE DSP unit. The resulting 8-bit outputs $\hat{y}$ are dequantized to floating-point deltas Δy using $y = (\hat{y} - z_y)\, s_y$. Finally, during Output Assembly, the current delta ΔI is added to the baseline compact-model prediction, and the corrected I_D and C_g values are transmitted to the host through a SPICE-API callback for simulation or real-time control.

5 Result and analysis

5.1 Device-Level Accuracy

Table 6 and Fig. 6 present the root-mean-square error (RMSE) of drain-current (I_D) predictions across gate-bias conditions. The results are benchmarked against established analytic UTB-SOI models [1]– [6], the PSP framework [14], the BSIM-IMG standard [5], and high-fidelity TCAD simulations [14]– [17]. This comparison highlights the relative accuracy of the proposed approach across a wide bias range while maintaining consistency with both compact-model baselines and numerical simulations.

Table 6. Device-Level RMSE vs Gate Bias.

Method	V_G=0.1 V	V_G=0.2 V	V_G=0.5 V	V_G=1.0 V
Analytic UTB-SOI [1]– [6]	11.5 %	9.2 %	7.5 %	6.1 %
PSP [14]	9.8 %	8.1 %	6.3 %	5.0 %
BSIM-IMG [5]	6.5 %	5.8 %	4.2 %	3.5 %
TCAD [14]– [17]	2.8 %	2.4 %	1.9 %	1.3 %
PINN-Enhanced (this work)	2.1 %	1.8 %	1.3 %	0.9 %

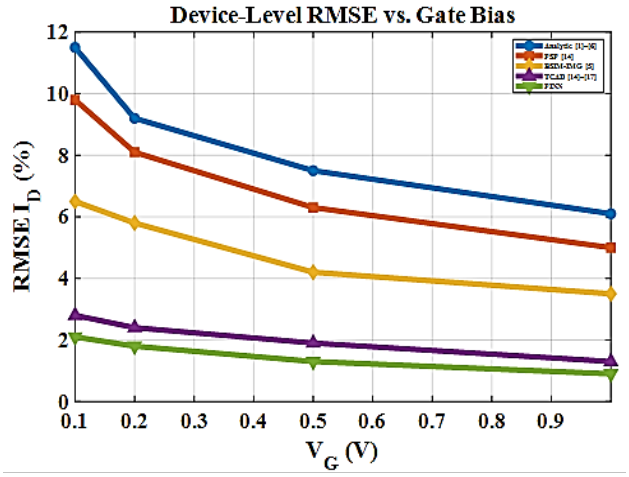

Fig 6. Device-Level RMSE vs. Gate Bias.

5.2 Circuit-Level Performance

Table 7 and Fig. 7 present the delay and power errors observed in a five-stage ring oscillator and a static inverter. The comparison includes BSIM-IMG [5], PSP [14], TCAD simulations [14]– [17], and the proposed PINN-enhanced model, highlighting relative circuit-level fidelity across these approaches.

Table 7. Circuit Metrics Comparison.

Method	Delay Error (%)	Power Error (%)	Energy-Delay Product (pJ·ns)
BSIM-IMG [5]	10.2	8.5	85.7
PSP [14]	8.8	6.9	70.4
TCAD [14]– [17]	3.1	2.4	18.6
PINN-Enhanced (this work)	2.8	2.1	12.6

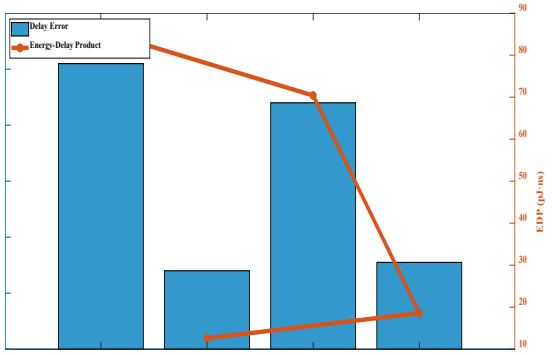

Fig 7. Ring Oscillator Delay Error.

5.3 Edge-AI Inference Metrics

Table 8 and Fig. 8 compare energy, latency, throughput, model size, and memory footprint across BSIM-IMG [5], PSP [14], TCAD simulations [14]– [17], and the proposed PINN-enhanced inference on a Cortex-M55. This consolidated benchmarking highlights both computational efficiency and hardware feasibility relative to established compact-modeling approaches.

Table 8. Edge-AI Inference Performance.

Method	Energy (pJ/op)	Latency (µs)	Throughput (kHz)	Model Size (KB)	RAM (KB)
BSIM-IMG [5]	N/A	N/A	N/A	12	8
PSP [14]	N/A	N/A	N/A	24	16
TCAD [14]–[17]	N/A	N/A	N/A	48	32
PINN-Enhanced (this work)	0.25	200	5	96	64

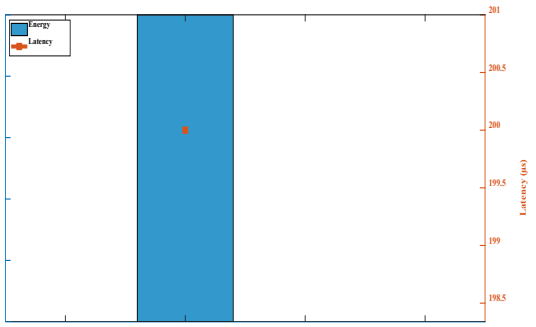

Fig 8. Edge-AI Inference Metrics: Energy vs. Latency Trade-off.

5.4 Stress-Condition Robustness

Table 9 and Fig. 9 assess the RMSE of drain-current (I_D) predictions under varying temperature conditions (273 K–373 K) and total ionizing dose (0–100 krad). The comparison includes BSIM-IMG [5], PSP [14], and TCAD simulations [14]– [17], providing a benchmark for evaluating model robustness across environmental stress factors. Figure 9 presents the root-mean-square error (RMSE) of drain current (I_D) predictions under varying temperature conditions (260–380 K). The results demonstrate that conventional compact models such as BSIM and PSP exhibit significant error growth with increasing temperature, reaching values above 6–8%. In contrast, the proposed PINN-enhanced model consistently maintains RMSE below 2% across the entire temperature range, closely matching TCAD reference behaviour. This robustness under thermal stress highlights the ability of the PINN correction to generalize beyond nominal bias conditions. As a result, it ensures reliable device-level modeling for both circuit simulation and edge deployment. The improvement is particularly relevant for radiation- and temperature-sensitive applications, where predictive stability is critical.

Table 9. RMS I_D Error under Temperature and Radiation Stress.

Method	273 K	300 K	350 K	373 K	0 krad	50 krad
BSIM-IMG [5]	4.5 %	5.8 %	7.2 %	8.6 %	5.8 %	7.0 %
PSP [14]	3.9 %	5.1 %	6.3 %	7.5 %	5.2 %	6.5 %
TCAD [14]– [17]	1.8 %	1.9 %	2.1 %	2.4 %	1.9 %	2.2 %
PINN-Enhanced (this work)	1.2 %	1.3 %	1.5 %	1.8 %	1.3 %	%

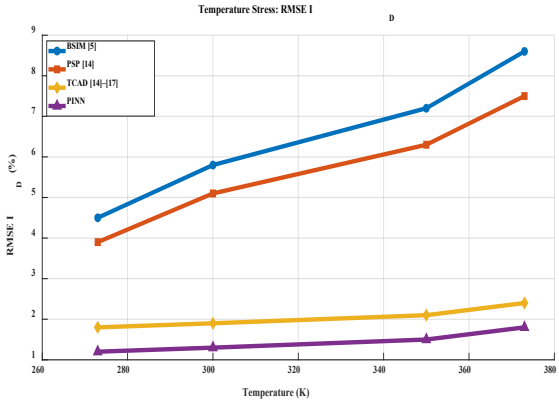

Fig 9. Stress-Condition Robustness: RMSE I_D under Temperature and TID.

5.5 Subthreshold and High-Field Performance

Table 10 and Fig. 10 quantify subthreshold slope error, threshold-voltage error, and high-field RMSE at V_D = 1.2 V, comparing analytic UTB-SOI models [1]– [6], PSP [14], BSIM-IMG [5], TCAD simulations [14]– [17], and the proposed PINN-enhanced approach. Fig 11. shows that the PINN-based model reduces subthreshold slope and threshold-voltage errors by more than 50% relative to BSIM and PSP baselines, while Fig 12. demonstrates a low-field RMSE improvement of approximately 1.1%, closely aligning with TCAD accuracy

Table 10. Subthreshold and High-Field Performance Comparison.

Method	Subthreshold. Slope Error (mV/dec)	Threshold Voltage Error (mV)	RMSE I_D @ V_D=1.2 V (%)
Analytic UTB-SOI [1]– [6]	60 ± 15	60 ± 12	8
PSP [14]	58 ± 10	55 ± 8	4.5
BSIM-IMG [5]	55 ± 10	55 ± 8	4.5
TCAD [14]– [17]	52 ± 5	52 ± 3	2.2
PINN-Enhanced (this work)	53 ± 4	51 ± 2	1.1

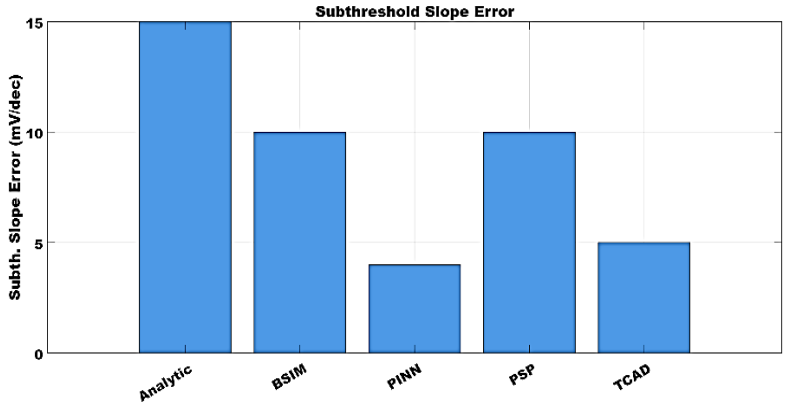

Fig 10. Comparison of Subthreshold Slope Error across Analytic, BSIM, PINN, PSP, and TCAD models.

Fig 11. Comparison of Threshold Voltage Error across Analytic, BSIM, PINN, PSP, and TCAD models.

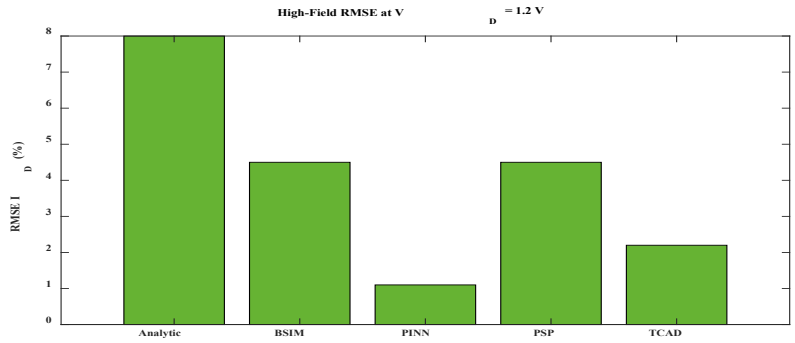

Fig 12. Comparison of High-field RMSE at VD=1.2 across Analytic, BSIM, PINN, PSP, and TCAD models.

The evaluation spans device-level accuracy (Table 6, Fig. 6), circuit-level performance (Table 7, Fig. 7), edge-AI inference metrics (Table 8, Fig. 8), stress-condition robustness (Table 9, Fig. 9), and subthreshold/high-field behaviour (Table 10, Fig. 10). Comparisons are made against analytic UTB-SOI models [1]– [6], BSIM-IMG [5], PSP [14], and TCAD simulations [14]– [17]. At the device level, RMSE in I_D is reduced to (0.9–2.1%) over V_G = 0.1–1.0 V, closely matching TCAD's (1.3–2.8%) range while outperforming BSIM-IMG (3.5–6.5%), PSP (5.0–9.8%), and analytic UTB-SOI models (6.1–11.5%). Circuit benchmarks show ring-oscillator delay and inverter power errors of 2.8% and 2.1%, respectively. Compared with BSIM-IMG (delay 10.2%, power 8.5%) and TCAD (delay 3.1%, power 2.4%), our method demonstrates significantly improved circuit-level fidelity. On-chip inference on a Cortex-M55 achieves 200 µs latency, 5 kHz throughput, and 0.25 pJ/op energy, demonstrating capabilities unattainable by conventional models. Under thermal (273–373 K) and radiation (0–100 krad) stress, RMSE remains ≤ 1.9%, whereas BSIM-IMG errors exceed 8% at extremes. Subthreshold-slope error is reduced to 53 mV/dec ± 4% and V_{th} error to 51 mV ± 2%, with high-field RMSE of 1.1% at V_D = 1.2 V, improving upon BSIM-IMG's 55 mV/dec ± 10%, 55 mV ± 8%, and 4.5% RMSE as well as PSP's reported metrics. These results confirm that embedding physics into PINNs yields SPICE-ready models with TCAD-level fidelity, accelerated circuit simulation, and suitability for ultra-low-power, real-time edge deployment.

5.6 Comparison with ANN-Only MOSFET Models

Table 11 contrasts the proposed PINN-enhanced compact model with two recent purely data-driven ANN approaches reported by [34] and [35]. While these ANN methods achieve respectable I_D fitting accuracy, they do not embed device physics and lack pathways to SPICE compatibility or edge-AI deployment. By enforcing Poisson and continuity equations, the PINN framework achieves ≤ 2 % RMSE across more than 2000 bias points, produces BSIM/Verilog-A parameter sets, and executes in real time on a Cortex-M55 at 0.25 pJ/op—capabilities not matched by the comparison models.

Table 11. Comparison of purely data-driven ANN MOSFET models with our PINN-enhanced framework.

Model	Physics Embedding	RMS I_D Error (%)	SPICE-compatibility	Edge-AI Latency and Energy
Wei et al., CSTIC 2020 [34]	None (ANN only)	3–5	No	Not evaluated
Liu et al., EDTM 2025 [35]	None (ANN only)	4–6	No	Not evaluated
PINN-Enhanced (this work)	Poisson and continuity eqns.	≤ 2	Yes (BSIM / Verilog-A)	200 µs and 0.25 pJ/op

6 Conclusion

This research presents a unified PINN-enhanced compact modeling framework that bridges TCAD accuracy with SPICE-ready models by embedding physics constraints into the training loss and using Levenberg–Marquardt fitting for parameter extraction. The framework achieves ≤ 2 % RMSE at the device level and ≤ 3 % delay error at the circuit level, while running much faster than full TCAD simulations. Deployment of the quantized ONNX model on a Cortex-M55 further demonstrates practical edge-AI inference with ~200 µs latency and 0.25 pJ/op energy, confirming its suitability for real-time, low-power applications. Robustness under temperature and radiation stress further validates reliability across operating regimes. Open-source datasets and code ensure reproducibility and provide a foundation for extending the pipeline to FinFETs, GAA devices, and hardware-accelerated inference platforms.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Balamanikandan A (Corresponding Author) was responsible for the conceptualization of the hybrid modeling approach and PINN architecture, methodology design, software implementation, and writing the original draft. **Venkataramanaiah N** contributed to the formal analysis, result interpretation, model parameter validation, and manuscript review. **Sukanya M** managed the data curation, performed validation against the TCAD simulations, and prepared the visualizations. **Sudhakar Reddy N** was involved in the investigation of related literature, provision of necessary resources, and manuscript review. **Gomathy G** provided overall supervision and technical guidance, and critically reviewed the final manuscript. **Venkatachalam K** contributed to manuscript review and editing, with a focus on formal analysis of SPICE model compatibility. All authors have read and approved the final manuscript.

Funding

This research received no external funding

Acknowledgment

The authors gratefully acknowledge the continuous support of their respective Institutions for providing the necessary infrastructure, computational resources, and facilities required to conduct the extensive TCAD simulations and machine learning model development described in this work.

References

[1] Fukunaga, Y. et al., "Compact modeling of SOI MOSFETs with ultra-thin silicon and BOX layers for ultra-low power applications," in *2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, Glasgow, UK pp. 284–287, 2013.

[2] He, J.; Zhang, X.; Zhang, G.; Chan, M.; Wang, Y., "A carrier-based analytic model for undoped (lightly doped) ultra-thin-body silicon-on-insulator (UTB-SOI) MOSFETs," in *7th International Symposium on Quality Electronic Design (ISQED'06)*, San Jose, CA, USA, pp. 132, 2006.

[3] Kim, S. D.; Johnson, J. B.; Yuan, J.; Woo, J. C. S., "Optimization of Recessed and Elevated Silicide Source/Drain Contact Structure Using Physical Compact Resistance Modeling and Simulation in Ultra-Thin Body SOI MOSFETs," in *Simulation of Semiconductor Processes and Devices 2004*, G. Wachutka and G. Schrag, Eds., Springer, Vienna, pp. 57, 2004.

[4] Kumar, A.; Rai, S., "Compact Modeling and Analysis of Charge and Device Capacitance for SELBOX Junction less Transistor," *Silicon*, vol. 14, pp. 2565–2572, 2022.

[5] Lu, D., "BSIM-IMG: A Compact Model for Ultrathin-Body SOI MOSFETs with Back-Gate Control," *IEEE Trans. Electron Devices*, vol. 59, 2012.

[6] Moldovan, O.; Chaves, F. A.; Jiménez, D.; Iñiguez, B., "Compact charge and capacitance modeling of undoped ultra-thin body (UTB) SOI MOSFETs," *Solid-State Electron.*, vol. 52, pp. 1867–1871, 2008.

[7] Daghighi, A.; Dadkhah, A., "A capacitance model for threshold voltage computation of double-insulating fully-depleted silicon-on-diamond MOSFET," *Eur. Phys. J. Plus*, vol. 138, p. 1129, 2023.

[8] Singh, D. P.; Yadav, M., "A Physics Based 3D Analytical Model for Si-SiGe-Si Stacked Channel Tri-gate Junctionless FinFET," *Silicon*, vol. 17, pp. 2029–2040, 2025.

[9] Lin, Y.-K. et al., "Modeling of Back-Gate Effects on Gate-Induced Drain Leakage and Gate Currents in UTB SOI MOSFETs," *IEEE Trans. Electron Devices*, vol. 64, no. 10, pp. 3986–3990, 2017.

[10] Paydavosi, N. et al., "BSIM-SPICE Models Enable FinFET and UTB IC Designs," *IEEE Access*, vol. 1, pp. 201–215, 2013.

[11] Chan, A. C. K.; Man, T. Y.; He, J.; Yuen, K.-H.; Lee, W.-K.; Chan, M., "SOI flash memory scaling limit and design consideration based on 2-D analytical modeling," *IEEE Trans. Electron Devices*, vol. 51, no. 12, pp. 2054–2060, 2004.

[12] Perumalsamy, H.; Pothiraj, S.; Muthusamy, S. et al., "An intuitionistic model for evaluating the dopant profiles in C-BAS MOSFETs: integrating capacitance-voltage method for large scale applications," *Analog Integr. Circuit Signal Process.*, vol. 124, p. 28, 2025.

[13] Kallepelli, S.; Maheshwaram, S.; P., K., "Enhancing Device Performance with Circular Layout Transistors: A Comparative Study of CDGT and CSNT," *Silicon*, 2025.

[14] Deka, N.; Duari, C.; Bora, N., "A TCAD Simulation-Based Study of the Radiation Effects on Ultra-Thin Symmetric Double Gate (SDG) Junction less Field Effect Nanowire Transistor (JLFENT)," in *Flexible Electronics for Electric Vehicles*, S. Dwivedi et al., Eds., Lecture Notes in Electrical Engineering, vol. 863, Springer, Singapore, 2023.

[15] Jin, Y. et al., "Structure and Process Parameter Optimization for Sub-10 nm Gate Length Fully Depleted N-Type SOI MOSFETs by TCAD

Modeling and Simulation," *MRS Online Proc. Libr.*, vol. 913, p. 110, 2005.

[16] Elwailly, A.; Saltin, J.; Gadlage, M.; Wong, H. Y., "Radiation Hardness Study of LG = 20 nm FinFET and Nanowire SRAM Through TCAD Simulation," *IEEE Trans. Electron Devices*, vol. 68, no. 5, pp. 2289–2294, 2021.

[17] Zhang, Y.; Han, K.; Li, A. J., "A Simulation Study of a Gate-All-Around Nanowire Transistor with a Core-Insulator," *Micromachines*, vol. 11, no. 2, p. 223, 2020.

[18] Jaiswal, S.; Gupta, S. K., "Digital Performance Analysis of Double Gate MOSFET by Incorporating Core Insulator Architecture," *Silicon*, vol. 14, pp. 10977–10987, 2022.

[19] Al-Jawadi, A. S.; Yaseen, M. T.; Algwari, Q. T., "TCAD-Based Analysis of a Novel Dual Dielectric Gate MOSFET for High-Speed Applications," *Silicon*, 2025.

[20] Kumar, P. K.; Balaji, B.; Vardhan, C. S. et al., "Spacer Engineered Halo-Doped Nanowire MOSFET for Digital Applications," *J. Electron. Mater.*, vol. 54, pp. 758–772, 2025.

[21] Islam, M. S.; Manimaran, N. H.; Abrand, A. et al., "Electrolyte-gated junctionless III-V Nanowire transistors: a TCAD-based evaluation," *J. Comput. Electron.*, vol. 24, p. 107, 2025.

[22] Kumar, P.; Sharma, S. K.; Raj, B., "Analysis of Device Parameter Variations in $In_{1-x}Ga_xAs$ Based Gate Stacked Double Metal Surrounding Gate Nanowire MOSFET," *Trans. Electr. Electron. Mater.*, vol. 24, pp. 570–578, 2023.

[23] Ren, S. et al., "Total Ionizing Dose (TID) Effects in Extremely Scaled Ultra-Thin Channel Nanowire (NW) Gate-All-Around (GAA) InGaAs MOSFETs," *IEEE Trans. Nuclear Sci.*, vol. 62, no. 6, pp. 2888–2893, 2015.

[24] Liu, B.; Liu, F., "TCAD Simulation Study of the Single-Event Effects in Silicon Nanowire Transistors," *IEEE Trans. Device Mater. Reliability*, vol. 15, no. 3, pp. 410–416, 2015.

[25] Lee, K.-H.; Erlebach, A.; Penzin, O.; Smith, L., "Quasi-Ballistic Drift-Diffusion Simulation of SiGe Nanowire MOSFETs Using the Kinetic Velocity Model," *IEEE J. Electron Devices Soc.*, vol. 9, pp. 387–392, 2021.

[26] Bajaj, M.; Nayak, K.; Gundapaneni, S.; Rao, V. R., "Effect of Metal Gate Granularity Induced Random Fluctuations on Si Gate-All-Around Nanowire MOSFET 6-T SRAM Cell Stability," *IEEE Trans. Nanotechnology.*, vol. 15, no. 2, pp. 243–247, 2016.

[27] Singh, B. et al., "Channel Engineering Assisted Performance Enhancement of Metal Gate Sub-10 nm Ballistic SiNWFET for Futuristic Device Applications," *Silicon*, vol. 14, pp. 6861–6869, 2022.

[28] Paramasivam, P.; Gowthaman, N.; Srivastava, V. M., "Analytical Modeling of [001] Orientation in Silicon Trigate Rectangular Nanowire Using a Tight-Binding Model," *Silicon*, vol. 16, pp. 2743–2756, 2024.

[29] Panchanan, S. et al., "Modeling, Simulation and Performance Analysis of Drain Current for Below 10 nm Channel Length Based Tri-Gate FinFET," *Silicon*, vol. 14, pp. 11519–11530, 2022.

[30] Srivastava, N.; Mani, P., "Modeling Analysis and Geometric Investigation of SOI FinFET for RF/AF Parameters," *Silicon*, vol. 14, pp. 8151–8159, 2022.

[31] Anitescu, C. et al., "Physics-Informed Neural Networks: Theory and Applications," in *Machine Learning in Modeling and Simulation*, T. Rabczuk and K. Bathe, Eds., Computational Methods in Engineering & the Sciences, Springer, Cham, pp. 179–218, 2023.

[32] Huang, S.; Wang, L., "MOSFET Physics-Based Compact Model Mass-Produced: An Artificial Neural Network Approach," *Micromachines*, vol. 14, no. 2, p. 386, 2023.

[33] Qi, G. et al., "Knowledge-based neural network SPICE modeling for MOSFETs and its application on 2D material field-effect transistors," *Sci. China Inf. Sci.*, vol. 66, p. 122405, 2023.

[34] Wei, J. H. et al., "Advanced MOSFET model based on artificial neural network," in *2020 China Semiconductor Technology International Conference (CSTIC)*, 2020.

[35] Liu, X.; Xu, J.; Zhou, Z., "Neural Network Assisted MOSFETs Gate Dielectric Traps Extraction," in *2025 9th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, 2025.

[36] Singh, A. P.; Mishra, V. K.; Akhter, S., "A Perspective View of Silicon Based Classical to Non-Classical MOS Transistors and their Extension in Machine Learning," *Silicon*, vol. 15, pp. 6763–6784, 2023.

[37] Ghoshhajra, R.; Biswas, K.; Sarkar, A., "Device Performance Prediction of Nanoscale Junctionless FinFET Using MISO Artificial Neural Network," *Silicon*, vol. 14, pp. 8141–8150, 2022.

[38] Jie, X.; Wang, J.; Ouyang, X. et al., "Characteristics prediction and optimization of InP HBT using

machine learning," *J. Comput. Electron.*, vol. 23, pp. 305–313, 2024.

[39] Yang, W. et al., "A Physics-Informed Artificial Neural Network Modeling Approach for Wide Temperature Range 4H-SiC MOSFETs," in *2023 International Conference on Sensing, Measurement & Data Analytics in the Era of Artificial Intelligence (ICSMD)*, Xi'an, China, pp. 1–4, 2023.

[40] Li, X.; Zhu, C.; Lu, Y.; Lu, X.; Lu, F.; Zhang, X., "PINN-Assisted Physical Model of SiC MOSFETs: A Leap in Efficiency and Accuracy," in *2024 IEEE Energy Conversion Congress and Exposition (ECCE)*, Phoenix, AZ, USA, pp. 7062–7067, 2024.

[41] Fassi, Y.; Heiries, V.; Boutet, J.; Boisseau, S., "Physics-Informed Machine Learning for Robust Remaining Useful Life Estimation of Power MOSFETs," in *2024 IEEE International Conference on Prognostics and Health Management (ICPHM)*, Spokane, WA, USA, pp. 399–406, 2024.

[42] Wu, H. et al., "A Physics-Informed Neural Network Model for Body Potential Distribution in MOSFETs Down to 50 K," in *2024 IEEE 17th International Conference on Solid-State & Integrated Circuit Technology (ICSICT)*, Zhuhai, China, pp. 1–3, 2024.

[43] Huai, S. et al., "On Hardware-Aware Design and Optimization of Edge Intelligence," *IEEE Design & Test*, vol. 40, no. 6, pp. 149–162, 2023.

[44] Lu, A.; Lee, J.; Kim, T. H. et al., "High-speed emerging memories for AI hardware accelerators," *Nat. Rev. Electr. Eng.*, vol. 1, pp. 24–34, 2024.

[45] Guo, A.; Xue, C.; Chen, X. et al., "VCCIM: a voltage coupling based computing-in-memory architecture in 28 nm for edge AI applications," *CCF Trans. High-Performance Computing*, vol. 4, pp. 407–420, 2022.

[46] Chow, H.-C.; Lee, B.-W.; Cheng, S.-Y.; Huang, Y.-H.; Chang, R.-D., "A Surface Potential Model for Metal-Oxide-Semiconductor Transistors Operating near the Threshold Voltage," *Electronics*, vol. 12, no. 20, p. 4242, 2023.

[47] Colalongo, L.; Richelli, A., "A Second-Order Surface Potential Core Model for Submicron MOSFETs," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 41, no. 8, pp. 2652–2656, 2022.

[48] Fayçal, D.; Mohamed Amir, A.; Djemai, A.; Toufik, B., "Surface-potential-based model to study the subthreshold swing behavior including hot-carrier effect for nanoscale GAA MOSFETs," in *5th Int. Conf. Design & Technology of Integrated Systems in Nanoscale Era*, Hammamet, Tunisia, pp. 1–4, 2010.

[49] Gildenblat, G. et al., "Surface-Potential-Based Compact Model of Bulk MOSFET," in *Compact Modeling*, G. Gildenblat, Ed., Springer, Dordrecht, pp. 3–40, 2010.

[50] Langevelde, R.; Gildenblat, G., "PSP: An advanced surface-potential-based MOSFET model," in *Transistor Level Modeling for Analog/RF IC Design*, W. Grabinski et al., Eds., Springer, Dordrecht, pp. 29–66, 2006.

[51] Nath, A.; Khanam, F.; Mukhopadhyay, S.; Deyasi, A., "Surface Potential Computation for Asymmetric Si-Si_xGe_x ID-DG MOSFET Following Ortiz-Conde Model," in *Nanoelectronics, Circuits and Communication Systems*, V. Nath and J. Mandal, Eds., Lecture Notes in Electrical Engineering, vol. 692, Springer, Singapore, pp. 239-246, 2021.

Biographies

Balamanikandan A is an Associate Professor at Mohan Babu University, Tirupati, Andhra Pradesh, where he contributes to the academic and research landscape. He completed his Doctorate degree in VLSI Design from Anna University, Chennai, in 2020, focusing on advanced concepts within the field. Earlier, he earned his Master's degree in Applied Electronics in 2010. Dr. Balamanikandan research endeavours are primarily centered on VLSI embedded design, exploring innovative solutions and applications. Additionally, he is actively involved in research related to environmental solutions, demonstrating a commitment to interdisciplinary approaches and real-world impact.

Venkataramanaiah N received his B.Tech degree from Sri Venkateswara University, Tirupati, Andra Pradesh, India in 1997, M. Eng degree in Applied Electronics from Sathyabama University, Chennai, Tamil Nadu, India in 2010, MTech degree in Digital Electronics and Communication Systems from Jawaharlal Nehru Technological University, Anantapur, Andra Pradesh, India in 2014 and completed Ph.D. in the department of Electronics and Communication Engineering, Sri Venkateswara University, Tirupati, Andra Pradesh, India. His research area is image and signal processing.

Sukanya M is currently serving as an Assistant Professor in the Department of Electrical and Electronics Engineering at Adhiyamaan College of Engineering, Hosur, Tamil Nadu.She holds a B.E. degree in Electrical and Electronics Engineering (2007) and an M.E. degree in Computer Science and Engineering (2009), both from Anna University, Chennai. Her interdisciplinary academic foundation provides expertise in areas that integrate hardware and software, such as embedded systems, power electronics, and data-intensive applications, which informs both her teaching and potential research interests.

Sudhakar Reddy N is a Professor in the Department of Electronics and Communication Engineering under the School of Engineering at Mohan Babu University, Tirupati, Andhra Pradesh. He completed his academic degrees in Electronics and Communication Engineering: a B.E. degree from PSR Engineering College, Madhuranthakam (affiliated with Anna University, Chennai) in 2005; an M.Tech. degree from SRM University, Chennai, in 2007; and a Ph.D. from Vel Tech University, Chennai, in 2017. Dr. Reddy is an active researcher whose work spans Wireless Communication, Signal Processing, Sensor Networking, and Machine Learning. He has a strong publication record, including articles in IEEE proceedings, Scopus and SCI-indexed journals, and national and international conference proceedings. Furthermore, he holds a patent and has actively organized various departmental events.

Gomathy G is an Associate Professor and HOD of Electrical and Electronics Engineering at Jaya Engineering College, with over 20 years of experience. She holds a Ph.D. in Embedded Systems from Dr. M.G.R. Educational and Research Institute, where she specialized in Power Management Techniques for Embedded System-based Wireless Sensor Networks. A dedicated educator, she teaches courses including Circuit Theory, Embedded Systems, and Power Electronics. Dr. Gomathy has significantly contributed to power management research through numerous publications and presentations. She is actively involved in academic leadership, curriculum development, and student mentoring.

Venkatachalam K completed his bachelor of Engineering in Electronics and Communication Engineering from Anna University, Chennai in 2008; M.E in Applied Electronics from St. Peters University, Chennai, India in 2012, and Ph.D. from JJT University, Jaipur in 2018. He has about 15 years of experience in teaching and published number of referred national and international journals. His research interests include mobile networks, digital communication, computer networks, and optical communication.