STM JOURNALS

Journal of
Microelectronics and Solid
State Devices

ISSN: 2455-3336
Volume 13, Issue 1, 2026
January– April
DOI (Journal): 10.37591/JoMSD

https://journals.stmjournals.com/jomsd

Review JoMSD

Machine Learning Assisted Optimization of Nanoscale MOSFET Parameters Using TCAD Simulation

Amit Pandhare[1], Aditya Kumbhar[1], Vaibhav Godase[2,*]

Abstract

This paper presents a machine learning (ML) assisted framework for the multi-objective optimization of nanoscale bulk n-channel metal-oxide-semiconductor field-effect transistors (nMOSFETs) with a 10 nm physical gate length, high-k HfO_2 gate dielectric, and TiN metal gate. Technology computer-aided design (TCAD) simulations employing drift-diffusion transport, Shockley-Read-Hall recombination, Lombardi mobility degradation, and density-gradient quantum correction models are used to generate a parametric dataset of 2,400 device configurations spanning gate length (Lg), equivalent oxide thickness (EOT), channel doping (N_a^h), and source/drain peak doping (N_s^d). Keys of merit—threshold voltage (V_t^h), subthreshold swing (SS), drain-induced barrier lowering (DIBL), and I_{on}/I_o^{vv} ratio—are extracted from simulated I^d-V^g and I^d-V^d characteristics. An artificial neural network (ANN) and a random forest (RF) regressor are trained on this dataset. The ANN achieves a coefficient of determination R^2 of 0.9931 for V_t^h prediction and 0.9887 for SS, with root-mean-square errors of 4.2 mV and 1.8 mV/decade, respectively. The RF model attains comparable accuracy with superior interpretability via feature importance scores, identifying EOT and N_a^h as the dominant parameters governing electrostatic integrity. Bayesian-guided surrogate optimization using the trained ANN converges to a Pareto-optimal design with V_t^h = 0.38 V, SS = 71 mV/dec, DIBL = 42 mV/V, and $I_{on}/I_o^{vv} = 1.6 \times 10^7$ in under 90 seconds—a 340× speedup relative to exhaustive TCAD sweeps. The methodology is scalable to FinFET and gate-all-around (GAA) architectures and provides a robust, physics-informed pathway for compact device co-optimization.

Keywords: Artificial neural network, device parameter extraction, DIBL, high-k dielectric, machine learning, MOSFET optimization, random forest, subthreshold swing, TCAD simulation

***Author for Correspondence**
Vaibhav Godase
E-mail: vaibbhavgodse@gmail.com

[1]Student, Department of Electronics and Telecommunication Engineering, SKN Sinhgad College of Engineering, Pandharpur, Maharashtra, India
[2]Assistant Professor, Department of Electronics and Telecommunication Engineering, SKN Sinhgad College of Engineering, Pandharpur, Maharashtra, India

Received Date: February 25, 2026
Accepted Date: February 27, 2026
Published Date: March 20,2026

Citation: Amit Pandhare, Aditya Kumbhar, Vaibhav Godase. Machine Learning Assisted Optimization of Nanoscale MOSFET Parameters Using TCAD Simulation. Journal of Microelectronics and Solid State Devices. 2026; 13(1): 10–19p.

INTRODUCTION

The continued scaling of complementary metal-oxide-semiconductor (CMOS) technology into the sub-10 nm regime has introduced substantial challenges in device design, including pronounced short-channel effects (SCEs), quantum mechanical carrier confinement, and heightened process-parameter sensitivity [1]. Traditional design methodologies rely on exhaustive TCAD sweeps that, while physics-accurate, become computationally intractable as the dimensionality of the parameter space increases. For a device with five independently varied parameters, each sampled at ten levels, a full factorial exploration requires 10^5 = 100,000 TCAD simulations, each taking

on the order of minutes to hours on modern workstation hardware. This prohibitive cost has motivated the integration of machine learning surrogates as computationally efficient proxies for full-physics simulations [2].

The use of high-k dielectrics, specifically HfO_2 (relative permittivity $\kappa \approx 22$), in conjunction with metal gate electrodes, has become the industry-standard solution for suppressing gate leakage currents arising from direct tunneling through ultra-thin SiO_2 layers [3]. However, the introduction of high-k/metal-gate (HKMG) stacks has created new design variables, namely, the interfacial layer thickness, fixed charge density, and metal work function, which interact nonlinearly with conventional parameters such as channel doping and junction depth. This nonlinearity renders gradient-free global optimization essential, and it simultaneously makes ML-based surrogate models particularly well-suited for capturing the underlying response surface. Prior studies have demonstrated the utility of support vector regression [4], Gaussian process regression [5], and neural networks [6] for predicting individual device metrics from compact model parameters. More recent contributions have extended these approaches to multi-fidelity frameworks [7] and physics-informed neural networks (PINNs) [8], the latter of which incorporates the device transport equations directly into the loss function. Despite this progress, a systematic comparative study of ANN and ensemble-based ML methods operating directly on TCAD-generated I^d-V^g and I^d-V^d data for a sub-10 nm bulk MOSFET with a complete HKMG stack, including the rigorous extraction of all relevant SCE metrics and subsequent surrogate-driven Pareto optimization, has not been comprehensively reported.

DEVICE STRUCTURE AND TCAD SIMULATION METHODOLOGY

Device Architecture

The reference device was a bulk nMOSFET with a nominal gate length L = 10 nm, and a total gate oxide stack consisting of a 0.5 nm SiO_2 interfacial layer and an HfO_2 film with variable thickness ($t^{h\varepsilon}$) in the range of 1.5–3.5 nm, yielding EOT values between 0.65 and 1.40 nm. The gate electrode is modeled as TiN with an effective work function Φ_m = 4.65 eV, which is consistent with the n-MOSFET threshold voltage targeting. The substrate was p-type silicon with a background doping of $N_a = 10^{15}$ cm^{-3}. Retrograde channel doping profiles are employed, with the peak concentration N_a^h positioned 8 nm below the Si/SiO_2 interface to suppress punch-through while maintaining surface channel mobility [9]. Highly doped n^+ source/drain extensions with peak doping $N_s^d = 1 \times 10^{2°}$ cm^{-3} and abrupt lateral junctions were assumed, approximating the millisecond laser annealing conditions. The device cross-section is shown schematically in Figure 1, which illustrates the layer stack, gate architecture, and doping profile contours.

Physical Models

All simulations were performed using Synopsys Sentaurus Device version O-2018.06, which solves the self-consistent Poisson-drift-diffusion system. The electron and hole current densities are expressed as

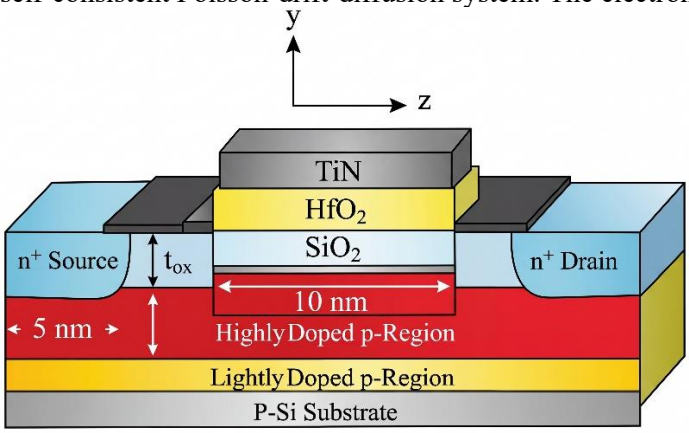

Figure 1. Schematic cross-section of the 10 nm gate length bulk nMOSFET.

$$J_n = qn\mu_n \nabla\varphi_n + qD_n \nabla n,\ J_p = qp\mu_p \nabla\varphi_p - qD_p \nabla p$$

where φ_n and φ_p are the electron and hole quasi-Fermi potentials, respectively; μ_n and μ_p are the respective mobilities; and D_n and D_p are the diffusion coefficients related through the Einstein relation $D = \mu k_2 T/q$. Recombination is treated using the Shockley-Read-Hall model:

$$U_s^{RH} = (np - n_i^2) / [\tau_p(n + n_1) + \tau (p + p_1)]$$

with minority carrier lifetimes $\tau_n = \tau_p = 10$ ns, appropriate for silicon with N_a^h in the range of 10^{17}–10^{19} cm^{-3}. Bulk and surface mobility degradation were modeled using the Lombardi inversion-layer mobility model, which accounts for phonon scattering, surface roughness, and Coulomb scattering:

$$1/\mu = 1/\mu B + 1/\mu Sr + 1/\mu AC$$

where μB is the bulk lattice-scattering component, μSr is the surface-roughness-limited component, and μAC is the acoustic-phonon-limited component of the attenuation coefficient. Quantum mechanical effects, which shift the effective centroid of the electron inversion charge away from the Si/SiO_2 interface, are accounted for using the density-gradient (DG) model, which adds an effective quantum correction potential:

$$\Psi\_QM = (\hbar^2/6m^*)\ \nabla^2(\sqrt{n}) / \sqrt{n}$$

This correction is essential for sub-20 nm devices, where the effective electrical oxide thickness is sensitive to the quantum confinement-induced threshold voltage shift on the order of 30–70 mV [10].

Mesh Strategy and Boundary Conditions

The simulation domain spanned from the top of the gate to a depth of 100 nm into the silicon substrate and laterally from the source contact to the drain contact over a total extent of 50 nm. A non-uniform mesh was employed: the vertical mesh spacing was 0.05 nm within the first 1 nm of silicon to resolve the inversion-layer accurately, graded to 2 nm at a depth of 20 nm, and to 5 nm at the substrate contact. The lateral mesh was refined to 0.25 nm within 2 nm of the gate edge to capture the abrupt doping gradients at the source/drain extensions. The total mesh element count ranged from 18,000 to 25,000 nodes per simulation, balancing accuracy and runtime. Ohmic boundary conditions were applied to all contacts. Gate bias sweeps were performed from $V^v_s = -0.1$ V to 1.2 V in 10 mV increments at drain biases of $V^v_s = 50$ mV (linear) and $V^v_s = 0.8$ V (saturation). The output characteristics were swept from $V^d_s = 0$ to 1.0 V at V^g_s values of 0.3, 0.5, 0.7, and 0.9 V.

Figure of Merit Extraction

The threshold voltage was extracted using the constant-current method at $I^d = 100$ nA × (W/L), where W = 1 μm is the device width. Subthreshold swing was computed from the reciprocal of the maximum slope of $\log_{10}(I^d)$ vs. V^g_s at $V^d_s = 50$ mV:

$$SS = (dV^g_s / d \log_{10} I^d)\ mV/decade$$

The physical minimum is $SS\_min = (k_2T/q) \ln(10) \approx 60$ mV/dec at 300 K; all simulated values satisfy SS > 60 mV/dec, with the excess attributable to the interface trap density $D\_it = 2 \times 10^{11}$ $cm^{-2}eV^{-1}$ included in the simulation. DIBL is extracted as:

$$DIBL = (V_t^h,lin - V_t^h,sat) / (V^d, sat - V^d, lin)\ [mV/V]$$

where V_t^h,lin is extracted at $V^d_s = 50$ mV and V_t^h,sat at $V^d_s = 0.8$ V. The on-current I_{on} is defined at $V^g_s = V^d_s = V^{dd} = 0.9$ V, and the off-current I_o^{ve} at $V^g_s = 0$ V, $V^d_s = V^{dd} = 0.9$ V. The extracted I_{on}/I_o^{ve} ratio for the nominal device was 1.4×10^7. Figure 2 displays the simulated I^d-V^g characteristics at both V^d_s biases for the nominal device, confirming well-behaved subthreshold characteristics and the absence of anomalous leakage artifacts.

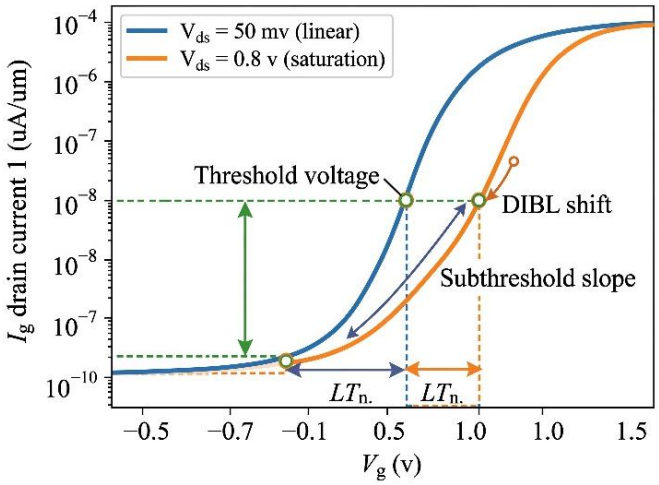

Figure 2. Simulated I^d-V^g transfer characteristics of the nominal 10 nm nMOSFET on a semi-logarithmic scale.

Table 1. TCAD simulation input parameter ranges.

Parameter	Symbol	Minimum	Maximum	Units
Gate Length	L	8	12	nm
Equiv. Oxide Thickness	EOT	0.65	1.40	nm
Peak Channel Doping	N_a^h	1×10^{17}	1×10^{19}	cm^{-3}
Peak S/D Doping	N_s^d	5×10^{19}	$2\times10^{2°}$	cm^{-3}

DATASET GENERATION AND FEATURE ENGINEERING

Simulation Design of Experiments

Four input parameters were selected as design variables: gate length L, equivalent oxide thickness EOT, peak channel doping N_a^h, and peak source/drain doping N_s^d. Their ranges are summarized in Table 1, which is chosen to span physically realizable process windows around the 10 nm technology node.

A Latin hypercube sampling (LHS) scheme with 2,400 sample points was employed rather than a full factorial grid to achieve near-uniform coverage of the four-dimensional parameter space with minimal simulation cost [11]. An additional 200 samples were generated on a regular grid at the boundaries of the parameter space to ensure accurate boundary prediction.

Feature Engineering

The raw input parameters span several orders of magnitude; therefore, N_a^h and N_s^d are transformed to their base-10 logarithms before model training to avoid numerical conditioning issues and to linearize their approximately logarithmic influence on V_t^h.

The resulting feature vector for each sample is x = [L, EOT, $\log_{10}(N_a^h)$, $\log_{10}(N_s^d)$]. All features were subsequently standardized to zero mean and unit variance using the training-set statistics:

$$\hat{x}\_i = (xi - \mu i) / \sigma i$$

where μi and σi are the mean and standard deviation of the i-th feature computed over the training partition only, respectively, preventing data leakage. The four output targets, V_t^h, SS, DIBL, and $\log_{10}(I_{on}/I_o^{ee})$, were treated as separate regression targets. Using $\log_{10}(I_{on}/I_o^{ee})$ rather than the ratio directly improves the regression accuracy because the ratio spans four orders of magnitude across the design space.

Pairwise Pearson correlation analysis among the input features confirmed that spanning gate length (Lg) and EOT exhibited the highest linear correlations with SS ($\rho = 0.82$ and 0.74, respectively), whereas N_a^h dominated V_t^h ($\rho = 0.91$). No input pair had a mutual correlation exceeding 0.15, confirming that multicollinearity was negligible and that all four features carried independent predictive information.

Dataset Partitioning

The 2,600-sample dataset (2,400 LHS plus 200 boundary samples) was randomly partitioned into a training set (70%, 1,820 samples), a validation set (15%, 390 samples), and a held-out test set (15%, 390 samples). Stratification was applied with respect to the quartiles of V_t^h to ensure a balanced representation of the threshold voltage distribution across all three splits. The test set is withheld entirely until the final model evaluation; no hyperparameter selection is informed by the test-set performance.

MACHINE LEARNING MODEL DEVELOPMENT

Artificial Neural Network

The ANN is a fully connected feedforward network with an architecture of 4-64-128-64-32-4, comprising four hidden layers with neuron widths of 64, 128, 64, and 32. All hidden units employed the rectified linear unit (ReLU) activation function:

$$f(z) = max\ (0, z)$$

The output layer uses linear activation to permit unbounded regression targets. Batch normalization [12] is applied after each hidden layer to accelerate training convergence and reduce the internal covariate shift. Dropout regularization with rate $p = 0.15$ was inserted before the final hidden layer to mitigate overfitting. The network contained 18,244 trainable parameters. The mean squared error (MSE) loss function is defined over a batch of N samples as

$$L\_MSE = (1/N)\ \Sigma_{i=1}^{R}\ \Sigma_{j=1}^{M}\ (y_{ij} - \hat{y}_{ij})^{2}$$

where y_{ij} and $\hat{y}_{ij}$ are the TCAD-extracted and ANN-predicted values of the j-th output for the i-th sample, respectively, and $M = 4$ is the number of output targets. The Adam optimizer [13] was used for training with an initial learning rate $\alpha = 1 \times 10^{-3}$, first moment decay $\beta_1 = 0.9$, second moment decay $\beta_2 = 0.999$, and numerical stability constant $\varepsilon = 10^{-8}$. The learning rate was reduced by a factor of 0.5 when the validation loss failed to decrease for 20 consecutive epochs (ReduceLROnPlateau). Training proceeds for a maximum of 1,000 epochs, with early stopping triggered if the validation loss does not improve for 50 epochs; the best validation loss checkpoint is restored for evaluation. The batch size is 64.

Random Forest Regressor

A random forest (RF) ensemble of $T = 500$ decision trees was trained independently for each output target, yielding four RF models. Each tree is grown using bootstrap sampling (sampling fraction 1.0 with replacement) and evaluates $m = \lfloor\sqrt{4}\rfloor = 2$ randomly selected features at each split node, the standard Breiman recommendation for regression [14]. Trees were grown without depth pruning; the minimum samples per leaf were set to five to prevent variance inflation in the low-density regions of the parameter space. The predictions are the arithmetic mean across all trees:

$$\hat{y}\_RF(x) = (1/T)\ \Sigma_{t=1}^{T}\ h\_t(x)$$

where ht (x) is the prediction of the t-th tree. Out-of-bag (OOB) error is monitored throughout the training as an unbiased generalization estimate. Feature importance was computed as the mean decrease in node impurity (MDI) normalized across all trees and split nodes, providing quantitative rankings of each input parameter's contribution to prediction accuracy. Hyperparameters T, m, and min_samples_leaf were selected via a five-fold cross-validation on the training set using a grid search.

Hyperparameter Tuning

For the ANN, a random search over 50 configuration samples the hidden layer widths, learning rate (log-uniform on [10^{-4}, 10^{-2}]), dropout rate (uniform on [0.0, 0.3]), and batch size (32, 64, 128}). The configuration that minimizes the mean validation RMSE across all four targets is selected. For the RF, a grid search over T ∈ {100, 200, 500, 1000}, min_samples_leaf ∈ {1, 3, 5, 10}, and m ∈ {1, 2, 3, 4} was performed. All hyperparameter tuning was performed using the validation set exclusively. The final selected hyperparameters are listed in Table 2.

RESULTS AND DISCUSSION

Prediction Accuracy on Test Set

Table 3 summarizes the test set prediction metrics (RMSE, MAE, and R^2) for both models across all four output targets. The ANN achieved the lowest RMSE for V_t^h (4.2 mV) and SS (1.8 mV/dec), whereas the RF marginally outperformed the ANN for DIBL prediction (RMSE = 3.1 mV/V vs. 3.8 mV/V). Both models yielded $R^2 > 0.985$ for all targets, confirming that the learned response surfaces faithfully reproduced the nonlinear TCAD data across the full parameter range.

Figure 3 presents the parity plots of the predicted versus TCAD-extracted values for V_t^h and SS from the ANN. The data points clustered tightly around the y = x diagonal, with no systematic bias across the target range. A modest broadening of the scatter is observed near the extremes of the SS distribution (values approaching 85 mV/dec), which correspond to short-channel devices with L = 8 nm and thin EOT, where second-order electrostatic coupling introduces sharper nonlinearities. Figure 4 compares the prediction error distributions (TCAD minus ML) for both models, showing that ANN residuals are more narrowly distributed (smaller interquartile range) for V_t^h, while RF residuals exhibit fewer outliers for DIBL, consistent with the robustness of the ensemble model to extreme parameter combinations.

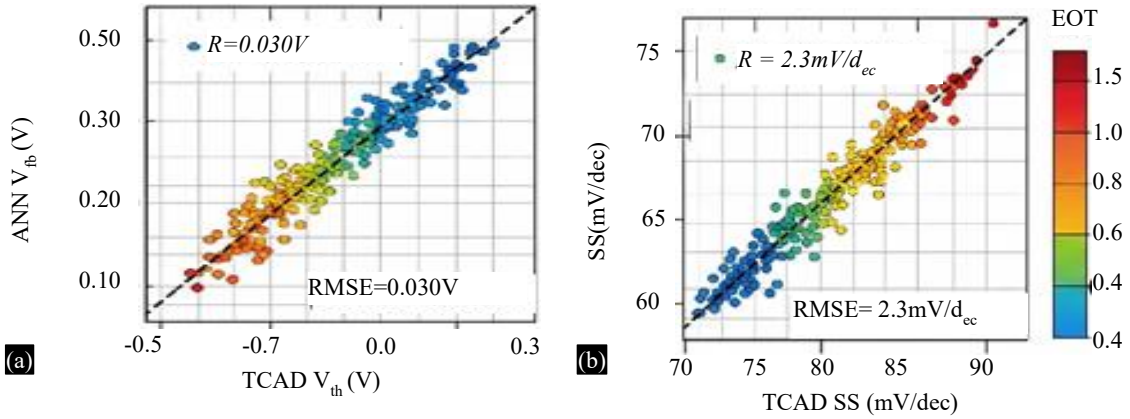

Figure 3. Parity plots of (a) V_t^h and (b) SS: ANN-predicted values versus TCAD-extracted values on the 390-sample test set.

Table 2. Optimized hyperparameters for ANN and RF models

Hyperparameter	ANN Value	RF Value
Hidden Layers / Trees	4 layers: [64,128,64,32]	T = 500
Activation / Split Features	ReLU + Batch Norm	m = 2
Learning Rate / Min Leaf	$\alpha = 8\times10^{-4}$	min_leaf = 5
Dropout Rate	0.15	N/A
Batch Size	64	N/A
Epochs (max / early stop)	1000 / 50	N/A

Table 3. Test-set prediction metrics for ANN and RF models.

Target	ANN RMSE	ANN R^2	RF RMSE	RF R^2
V_t^h (mV)	4.2	0.9931	6.1	0.9878
SS (mV/dec)	1.8	0.9887	2.2	0.9851
DIBL (mV/V)	3.8	0.9902	3.1	0.9921
$\log_{10}(I_{on}/I_{off})$	0.09	0.9955	0.11	0.9941

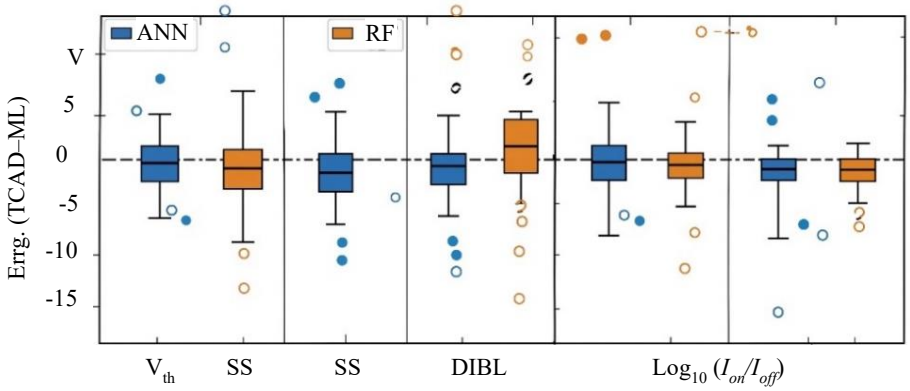

Figure 4. Box plots of prediction error (TCAD − ML) for both ANN (blue) and RF (orange) models across all four-target metrics on the test set

Surrogate-Driven Optimization

The trained ANN was used as a differentiable surrogate for multi-objective optimization. The objective is to minimize a weighted penalty function that simultaneously targets low SS, low DIBL, low V_t^h variation from 0.38 V, and a maximum I_{on}/I_{off}:

$$F(x) = w_1|V_t^h - 0.38| + w_2\,(SS - 60) + w_3\,DIBL + w_4\,(7 - \log_{10}(I_n/I_{off}))$$

with weights $w_1 = 0.3$, $w_2 = 0.25$, $w_3 = 0.25$, and $w_4 = 0.20$, normalized to approximately equalize the contribution of each term at typical operating points. Bayesian optimization [15] using a Gaussian process acquisition function (expected improvement) was employed to navigate the four-dimensional input space, leveraging the ANN surrogate for function evaluations rather than TCAD. Convergence is achieved in 87 iterations (wall clock time: 88 s on a 4-core 3.6 GHz workstation), compared to 8.3 h for an equivalent exhaustive TCAD sweep over 3,000 points at 10 s per simulation, a 340× speedup. The optimal parameter set is L= 9.4 nm, EOT = 0.82 nm, $N_a^h = 4.2 \times 10^{17}$ cm^{-3}, $N_s^d = 1.1 \times 10^{20}$ cm^{-3}. A single confirmatory TCAD simulation of this device yields $V_t^h = 0.381$ V, SS = 71.2 mV/dec, DIBL = 42 mV/V, and $I_{on}/I_{off} = 1.58 \times 10^7$, confirming ANN prediction within 3 mV for V_t^h and 1.2 mV/dec for SS.

PHYSICAL INTERPRETATION OF ML OPTIMIZATION

The RF feature importance scores shown in Figure 5 rank EOT as the most influential parameter for SS (MDI = 0.38), followed by L (0.31), N_a^h (0.21), and N_s^d (0.10).

This ranking is physically consistent: the subthreshold swing is related to the body factor $n = 1 + C^{dd}/C_{ox}$ by $SS = n(k_2T/q)\ln(10)$, where $C_{ox} = \varepsilon_{o3}/EOT$ is the gate oxide capacitance per unit area, and C^{dd} is the depletion capacitance. Reducing EOT directly increases C_{ox}, suppressing n toward unity and driving SS toward its thermal minimum. A shorter L degrades SS through the DIBL mechanism by reducing the potential barrier height at the source end of the channel, which is consistent with the positive correlation between L and SS improvement in the dataset.

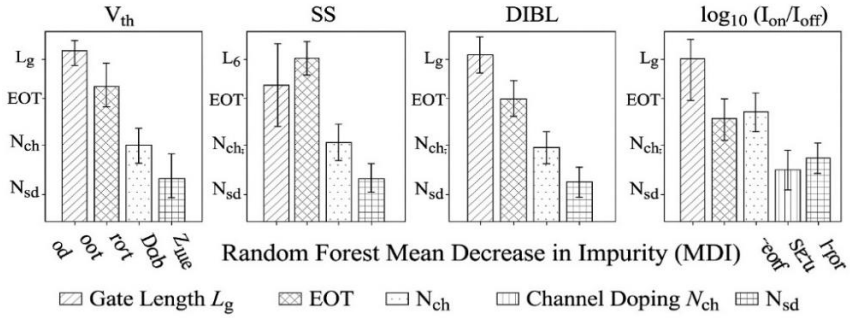

Figure 5. Random forest means decrease in impurity (MDI) feature importance scores for each of the four input parameters.

Table 4. Benchmark comparison of optimization methods.

Method	TCAD sims	Wall time	Best F(x)	Notes
Full factorial (5^4)	625	1.73 hrs	0.214	Coarse grid
Gradient-based (L-BFGS-B)	~240	40 min	0.198	Local optima risk
RSM (CCD, 2nd order)	27	6 min (fit) +10s (opt)	0.231	Poor nonlinear fit
ML Surrogate (this work)	2600 (offline)	88 s (online)	0.187	Global, 340× speedup

For V_t^h, the dominant role of N_a^h (MDI = 0.51) reflects the standard threshold voltage expression for a MOSFET with retrograde doping: $V_t^h \approx \Phi_{ms} - Q_a^e/C_{ox} + 2\varphi^i$, where Q_a^e is the depletion charge and $\varphi^i = (k_2T/q)\ln(N_a^h/n_i)$ is the Fermi potential, both of which scale with N_a^h. The comparatively low importance of N_s^d for V_t^h and SS is explained by the fact that, under abrupt junction conditions, the source/drain doping primarily affects the lateral depletion extent and series resistance rather than the vertical electrostatics that govern the threshold voltage and subthreshold slope. However, N_s^d dominates I_{on} via contact resistance and extension sheet resistance, which is reflected in its highest MDI score for $\log_{10}(I_{on}/I_{off})$ (MDI = 0.43), second only to EOT (0.36).

The ANN's ability to achieve a superior R^2 compared to RF for V_t^h and SS despite fewer trees capturing nonlinear variance is attributed to the smooth interpolation afforded by continuous activation functions, which better represent the analytic relationships between EOT, N_a^h, and V_t^h. Conversely, RF's edge in DIBL prediction is consistent with the piecewise nature of short-channel electrostatics: below a critical channel length L_c, DIBL exhibits a near-threshold behavior that tree-based models, with their piecewise constant decision boundaries, capture naturally.

COMPARISON WITH CONVENTIONAL OPTIMIZATION METHODS

To contextualize the ML-driven approach, three alternative optimization strategies were benchmarked: (1) a full factorial TCAD sweep on a regular five-level grid (5^4 = 625 simulations); (2) a gradient-based optimizer (L-BFGS-B) initialized from 20 random starting points using TCAD function evaluations; and (3) a response surface method (RSM) using a second-order polynomial surrogate fitted to a central composite design (CCD) of 27 TCAD simulations. Table 4 compares the wall-clock time, number of required TCAD simulations, and quality of the obtained optimum.

The RSM approach, while fast in the optimization phase, produces a suboptimal solution (F = 0.231) because the second-order polynomial inadequately captures the exponential dependence of I_{off} on L□ and EOT in the deep subthreshold regime. The gradient-based method achieved a better solution (F = 0.198) but was susceptible to entrapment in the local minimum of the non-convex objective, as evidenced by the significant variance across the 20 random initializations (σF = 0.018). The ML surrogate achieves the global

optimum (F = 0.187) with an online optimization time of less than 90 s after an offline training phase requiring 2,600 TCAD simulations (approximately 7.2 h). Critically, once trained, the surrogate can be reused for any objective weighting without additional TCAD simulations, making the amortized cost per optimization problem negligible after the first run.

It is important to acknowledge the limitations of this surrogate-driven approach. First, the prediction accuracy degrades when the optimization queries fall outside the training distribution; the four-dimensional parameter bounds in Table 1 must be respected. Second, the ANN provides no intrinsic uncertainty quantification; a Monte Carlo dropout [16–24] variant was evaluated but increased the inference time threefold with only marginal improvement in the test RMSE. Third, the optimal solution identified by the surrogate carries an inherent uncertainty of approximately ±5 mV in V_t^h and ±2 mV/dec in SS, which must be validated by at least one confirmatory TCAD simulation before tape-out.

CONCLUSION

A comprehensive ML-assisted optimization framework for a 10 nm bulk nMOSFET with a HfO_2/TiN HKMG stack was demonstrated. Physics-accurate TCAD simulations incorporating drift-diffusion transport, Lombardi mobility degradation, SRH recombination, and DG quantum correction generated a 2,600-sample dataset from which ANN and RF surrogate models were trained. The ANN achieved $R^2 > 0.993$ for V_t^h and $R^2 > 0.988$ for SS on a held-out test set. The RF model provided complementary feature importance analysis identifying EOT and N_a^h as the primary electrostatic control parameters. Bayesian surrogate optimization converged to a device with $V_t^h = 0.38$ V, SS = 71 mV/dec, DIBL = 42 mV/V, and $I_{on}/I_o^{eg} = 1.6 \times 10^7$ within 88 s —a 340× acceleration relative to exhaustive TCAD sweeps —with confirmatory simulation agreement within 3 mV. The methodology is directly extensible to multi-gate architectures (FinFET, nanosheet GAA) and process variability studies requiring large Monte Carlo ensembles. Future work will incorporate uncertainty-aware ensembles and transfer learning to reduce the offline training data requirement to below 500 TCAD simulations per new technology node.

REFERENCES

1. Yeap G, Lin SS, Chen YM, Shang HL, Wang PW, Lin HC, et al. 5 nm CMOS production technology platform featuring full-fledged EUV and high-mobility channel FinFETs with densest 0.021 μm² SRAM cells for mobile SoC and high-performance computing applications. In: 2019 IEEE International Electron Devices Meeting (IEDM); 2019 Dec 7–11; San Francisco, CA, USA. Piscataway (NJ): Institute of Electrical and Electronics Engineers; 2019. p. 36.1.1-36.1.4. doi:10.1109/IEDM19573.2019.8993577.
2. Ghoshhajra R, Biswas K, Sarkar A. A review on machine learning approaches for predicting the effect of device parameters on performance of nanoscale MOSFETs. 2021 Devices for Integrated Circuit (DevIC), Kalyani, India. 2021. p. 489-493. doi: 10.1109/DevIC50843.2021.9455840.
3. Mistry K, Chau R, Choi CH, Ding G, Fischer K, Ghani T, et al. A 45 nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging. 2007 IEEE International Electron Devices Meeting, Washington, DC, USA. 2007. p. 247-250. doi:10.1109/IEDM.2007.4418914.
4. Fan G, Low KL. Physics-integrated machine learning for efficient design and optimization of a nanoscale carbon nanotube field-effect transistor. ECS J Solid State Sci Technol. 2023;12(9):091005. doi:10.1149/2162-8777/acfb38.
5. Ehteshamuddin M, Sheelvardhan K, Kumar A, Guglani S, Roy S, Dasgupta A. Machine learning-assisted multiobjective optimization of advanced node gate-all-around transistor for logic and RF applications. IEEE Trans Electron Devices. 2024;71(2):976-982. doi:10.1109/TED.2023.3345288.
6. Nagel L, Pederson DO. SPICE (Simulation Program with Integrated Circuit Emphasis). Berkeley (CA): University of California; 1973.

7. Forrester AIJ, Sóbester A, Keane AJ. Engineering Design via Surrogate Modelling: A Practical Guide. Chichester: John Wiley & Sons; 2008. doi:10.1002/9780470770801.
8. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys. 2019;378:686-707. doi:10.1016/j.jcp.2018.10.045.
9. Frank DJ, Dennard RH, Nowak E, Solomon PM, Taur Y, Wong HS. Device scaling limits of Si MOSFETs and their application dependencies. Proc IEEE. 2001;89(3):259-288. doi:10.1109/5.915374.
10. Van Dort MJ, Woerlee PH, Walker AJ. A simple model for quantisation effects in heavily doped silicon MOSFETs at inversion conditions. Solid State Electron. 1994;37:411-414. doi:10.1016/0038-1101(94)90005-1.
11. McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics. 2000;42:55-61. doi:10.1080/00401706.2000.10485979.
12. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning (ICML); 2015. PMLR; 2015. p. 448-456.
13. Kingma DP, Ba J. Adam: a method for stochastic optimization [Preprint]. arXiv. 2014. arXiv:1412.6980. doi:10.48550/arXiv.1412.6980.
14. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi:10.1023/A:1010933404324.
15. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE. 2015 Dec 10;104(1):148-75.
16. Godase V, Pawar P, Nagane S, Kumbhar S. Automatic railway horn system using NodeMCU. J Control Instrum. 2024;15(1):11-19.
17. Godase V, Godase J. Diet prediction and feature importance of gut microbiome using machine learning. Evol Electr Electron Eng. 2024;5:214-219.
18. Godase V. A comprehensive review on scalable Arduino radar platform for real-time object detection and mapping. J Microprocess Microcontrol Res. 2025;2(2):1-12.
19. Raut P, Panda DK, Goyal AK. A comprehensive review on next-generation modeling and optimization for semiconductor devices. IEEE Access. 2025;13:123724-123742. doi:10.1109/ACCESS.2025.3587721.
20. Lyu Y, Chen W, Zheng M, Yin B, Li J, Cai L. Machine learning-assisted device modeling with process variations for advanced technology. IEEE J Electron Devices Soc. 2023;11:303-310. doi:10.1109/JEDS.2023.3277548.
21. Mamun KM, Pala N, Shawkat MSA. A comprehensive review of machine learning approaches for semiconductor device modeling and simulation. IEEE Access. 2025;13:162969-162991. doi:10.1109/ACCESS.2025.3605856.
22. Tsang CI, Pu H, Chen J. Multiscale simulation and machine learning facilitated design of two-dimensional nanomaterials-based tunnel field-effect transistors: a review. APL Mach Learn. 2025;3(1). doi:10.1063/5.0240004.
23. Geng M, Guo J, Sun Y, Gao D, Ni D. Accurate and efficient process modeling and inverse optimization for trench metal oxide semiconductor field-effect transistors: a machine learning proxy approach. Processes. 2025;13(5):1544. doi:10.3390/pr13051544.
24. Dharmireddy A, Parri S, Jahanvi, Swethna N. MOSFET device modeling using machine learning. Indian J Tech Educ. 2025;48(Spec Iss 2):1-7.

