Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/147/2025.22570

Lifetime Prediction of Semiconductor Devices Based on Deep Learning

Hongkai Weng

School of Mathematics (Zhuhai Campus), Sun Yat-sen University, Guangzhou, China
wenghk@mail2.sysu.edu.cn

Abstract: The reliability of power semiconductor devices, such as insulated gate bipolar transistors (IGBTs), is crucial for aerospace and industrial applications. Traditional prediction methods face challenges in multimodal data, integration physical constraints, and feature extraction accuracy. This study proposes a physics-informed deep learning framework for remaining useful life(RUL) prediction, by fusing high-frequency transient waveforms, steady-state thermal measurements, and electrical characterization data from source measurement units(SMUs). A hybrid architecture combines dilated convolutional neural networks (Dilated CNNs), to capture multi-scale transient features, long short-term memory (LSTM) networks with attention mechanisms for thermal sequence modeling and physics-guided loss functions incorporating the Coffin-Manson fatigue model. Experimental validation on NASAs accelerated aging dataset devices 2-5 demonstrated rapid convergence, with validation loss decreasing from 8362.9460 to 0.0224, and training loss from 0.4585 to 0.1121 over 100 epochs. The model achieved an RMSE of 0.0536 and an MAE of 0.0523, significantly outperforming non-dilated CNN baselines in convergence speed and stability.

Keywords: Deep learning, semiconductor devices, multimodal fusion, physics based neural networks, remaining useful life (RUL)

1. Introduction

Research on predicting remaining useful life (RUL) of power semiconductors primarily combines physical modeling and data driven methodologies [1]. The physical approach employs multi physics finite element simulations incorporating material fatigue equations to estimate degradation patterns [2], whereas data driven techniques utilize advanced analytics to identify trends within historical degradation data. While both methods demonstrate efficacy, they present complementary limitations that drive methodological innovations.

Neural network-enhanced data driven solutions now dominate recent advancements due to superior pattern recognition capabilities. Domestic researchers have implemented various architectures: GuoYuan Li's team [3] achieved junction temperature prediction using BP and RBF networks, while Cao Jiansheng's optimization of BP networks through particle swarm optimization (PSO) boosted nMOSFET lifetime estimation accuracy by 14.19% [4]. International studies demonstrate comparable progress-Kalman filter-integrated BP networks enable precise temperature tracking, and feedforward neural networks combined with principal component analysis attain over 97% diagnostic accuracy in inverter fault detection [5-7]. Comparatively, real time RUL prediction models now incorporate diverse machine learning techniques including BP networks, random forests,

DOI: 10.54254/2755-2721/147/2025.22570

and extreme learning machines [8]. This paper successfully addressed three key challenges: first, multi modal data fusion, which involved integrating 12.5 kHz transient waveforms, SMU parameters, and thermal sequences; second, physics aware modeling, where Coffin Manson fatigue equations were embedded into the loss functions; and third, robust validation, which entailed handling data inconsistencies such as missing transients and sensor drift.

This paper developed a hybrid deep learning framework that combines Dilated Convolutional Neural Networks (Dilated CNNs) to capture high frequency switching transients, Long Short-Term Memory networks (LSTMs) with attention mechanisms for thermal sequence modeling, and physics-guided loss terms to enforce degradation consistency. This paper provides the first integration of SMU characterization data (threshold voltage, leakage current) into RUL prediction, a deployable model achieving real time inference (<15 ms) on edge devices, and validation on NASA's accelerated aging datasets, demonstrating 99.8% validation loss reduction.

2. Data collection and preprocessing

2.1. Data sources

This study utilizes the accelerated aging dataset published by NASA Prognostics Center of Excellence [9], which covers degradation data of four sets of insulated gate bipolar transistors (IGBT, Devices 2-5) under thermal overstress. The dataset comprises the following multimodal information:

High frequency transient waveform: Sampling frequency of 12.5 kHz, acquisition period of 12500 sample points, including transient waveforms of gate emitter voltage (V_{ge}), collector emitter voltage (V_{ce}), and collector current (I_c). These waveforms record the voltage spikes and current oscillations caused by parasitic parameters during the switching process of the device, and are key indicators reflecting the dynamic characteristics of the device.

SMU static parameters: Threshold voltage (V_{th}), breakdown voltage (V_{br}), and leakage current (I_{leak}) obtained through the source measurement unit (SMU), covering 40 unaged MOSFET devices as reference.

Thermal sequence data: steady state measurement values of package temperature $(T_{package})$ and ambient temperature $(T_{ambient})$, used to evaluate the degradation trend of device heat dissipation performance.

2.2. Preprocessing pipeline

Data alignment and interpolation: To address missing transient waveform data caused by device communication interruptions, cubic spline interpolation is employed to reconstruct missing intervals. This approach preserves the nonlinear characteristics of high frequency signals and outperforms linear interpolation or polynomial fitting in maintaining signal integrity. Temperature calibration: To eliminate the influence of sensor drift, the steady state thermal data is aligned with the SMU measurement values using the least squares method. After calibration, the temperature error decreased from ± 2.1 ° C to ± 0.3 ° C.

Data augmentation: To mitigate sample imbalance issues, TimeGAN is introduced to generate synthesized waveforms. Through adversarial training, the generator learns the temporal distribution of real data and generates transient signals that are consistent with the statistical characteristics of the original data.

The Remaining Useful Life (RUL) is defined based on a failure criterion where the temperature exceeding threshold ($T_{package} > 330°C$), then RUL is as follows:

$$RUL = T_{failure} - T_{current} \quad (1)$$

DOI: 10.54254/2755-2721/147/2025.22570

where, $(T_{failure})$is the failure time, $(T_{current})$is the current time.

2.3. Feature engineering

To quantify device degradation under thermal electrical stress, three physics-based features were extracted from the raw multimodal data, as presented in Table 1. The Switching energy loss (E_{sw}) was calculated by integrating the product of collector emitter voltage $(V_{ce}(t))$ and collector current $(I_c(t))$ over each switching cycle. This metric directly reflects power dissipation during transient operations, where higher E_{sw} values correlate with accelerated bond wire fatigue due to joule heating. The Thermal resistance (R_{th}) was derived from the temperature gradient between the package $(T_{package})$ and ambient $(T_{ambient})$ normalized by steady state power loss $(P_{loss} = V_{ce} \cdot I_c)$. A rising R_{th} indicates deteriorating heat dissipation efficiency, often caused by delamination or solder joint cracks. Finally, the Threshold voltage shift (ΔV_{th}) was defined as the deviation of $V_{th}(t)$ from its initial value $(V_{th}(0))$, serving as a proxy for gate oxide degradation due to charge trapping.

These features were selected for their interpretability and alignment with known failure mechanisms in power semiconductors. For instance, E_{sw} captures dynamic stress during switching, while R_{th} and ΔV_{th} monitor gradual material degradation. To ensure compatibility across modalities, transient waveforms were downsampled to match the thermal sampling rate (1 Hz), and missing values in SMU parameters were imputed using adjacent cycle averages. The integration of these features enables the model to holistically address both abrupt and cumulative degradation patterns, forming a critical foundation for subsequent multimodal fusion.

Table 1: Extracted features and their physical significance

Feature	Formula	Purpose
E_{sw}	$\int V_{ce}(t) \cdot I_c(t)dt$	Switching energy loss
R_{th}	$(T_{package} - T_{ambient})/P_{loss}$	Degradation of heat dissipation performance
V_{th}	$\Delta V_{th} = V_{th}(t) - V_{th}(0)$	Gate oxide degradation

3. Multi modal deep learning architecture

3.1. Network design

The core algorithm architecture of this study combines dilated convolutional neural network (Dilated CNN), long short-term memory network (LSTM), and physical constraint loss function to achieve high precision life prediction through multimodal fusion.

3.2. Dilated CNN

The algorithmic characteristics of Convolutional Neural Networks (CNNs) make them uniquely valuable in the field of industrial inspection. In response to the anomaly detection requirements in semiconductor manufacturing, this network architecture can perform deep feature mining on high dimensional sensor data. Its hierarchical processing mechanism consists of three key modules: the feature extraction unit uses a sliding kernel function to model the spatial correlation of input data, and captures geometric differences such as micro cracks or circuit texture anomalies on the wafer surface through multi scale filters; The feature compression layer adopts a nonlinear dimensionality reduction strategy, which preserves effective information while suppressing data redundancy. Typical methods improve the model's noise resistance by preserving regional extremum or aggregating mean values; The decision-making module establishes classification boundaries through global feature association

DOI: 10.54254/2755-2721/147/2025.22570

and maps abstract features to the probability distribution of fault types. Dilated CNN is an improved convolutional neural network that introduces dilated convolution to expand the receptive field while maintaining the same number of parameters and computational complexity.

Dilated Convolution Principle:

By inserting holes (gaps) between convolutional kernel elements, the receptive field is expanded without increasing the number of parameters.

Mathematical expression:

$$y[i] = \sum_{k=1}^{K} x[i + d \cdot k] \cdot w[k] \tag{2}$$

where x is the input signal, y is the output signal, w is the convolution kernel, d is the dilation rate (representing the interval between convolution kernel elements), and K is the size of the convolution kernel.

3.3. LSTM

In the field of temporal data analysis, recurrent neural networks (RNNs) provide innovative solutions for industrial equipment status monitoring with their unique data modeling capabilities. In response to the reliability assessment requirements of semiconductor components, this architecture can model the timing of dynamic monitoring signals generated during equipment operation, and its core mechanism is to construct a computing unit with memory function. By introducing a hidden state transmission mechanism, the network can transmit the characteristic information of historical operating parameters to the current computing node, thereby achieving continuous analysis of the degradation trajectory of device performance.

However, the original loop architecture has shown significant limitations in practical applications. When faced with the analysis of device operation logs for several months, the network is prone to the phenomenon of weakened cross time feature correlation, which is due to the information dissipation effect in the process of error backpropagation. To address this bottleneck, the enhanced architecture incorporates a dynamic information regulation mechanism. The Long Short-Term Memory Network (LSTM) utilizes cell state channels in conjunction with a tripartite gating system—comprising the input gate, forget gate, and output gate—to dynamically manage the flow of information. The input gate selectively incorporates new operational data patterns, the forget gate discards obsolete or redundant noise, and the output gate modulates feature delivery to subsequent layers. This gated mechanism enables active filtering of non-essential signals while preventing gradient attenuation across extended time horizons.

The core formulas are as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(forget gate)} \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(input gate)} \tag{4}$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(candidate memory)} \tag{5}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \quad \text{(update memory)} \tag{6}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(output gate)} \tag{7}$$

$$h_t = o_t \odot \tanh(C_t) \tag{8}$$

3.4. Physical constraint loss function

Coffin-Manson model is adopted to describe the fatigue life caused by thermal cycling, which is expressed as:

DOI: 10.54254/2755-2721/147/2025.22570

$$N_f = A \cdot (\Delta T)^{\alpha} \tag{9}$$

where $A = 0.05, \alpha = -2.5$is the material constant, $\Delta T = T_{max} - T_{min}$ then align the $N_f$predicted by the physical model with the RUL predicted by deep learning:

$$\mathcal{L} = 0.7 \cdot \mathrm{MSE}(RUL_{pred}, RUL_{true}) + 0.3 \cdot |RUL_{pred} - N_f| \tag{10}$$

3.5. Cross-attention mechanism

Cross-attention is a variant of the attention mechanism designed to model relationships between two distinct sequences or modalities. Its core idea is to dynamically allocate attention weights by computing the relevance between a Query sequence and a Key-Value pair sequence, enabling effective fusion of information from different sources. Cross-attention involves three steps: similarity computation, weight allocation, and weighted summation.

1. Similarity Computation:

$$\text{Attention Scores} = \frac{QK^T}{\sqrt{d_k}} \tag{11}$$

where,

Query (Q): Features from one modality (here is transient waveform data).

Key (K) and Value (V): Features from another modality (here is thermal sequences or SMU parameters).

-Scaling factor ($\sqrt{d_k}$): prevents gradient explosion.

2. Weight Allocation:

$$\text{Attention Weights} = \text{Softmax}(\text{Attention Scores})$$

Softmax ensures weights sum to 1 for each row.

3. Weighted Summation:

$$\text{Output} = \text{Attention Weights} \cdot V$$

Output is a context-aware fusion of Value vectors.

4. Experimental results

4.1. Training dynamics

Validation loss reduction: The values converge rapidly, as evidenced by the decrease from 8362.9460 (Epoch 1) to 0.0224 (Epoch 100), as illustrated in Figure 2.

Training loss: The model stabilized at a value of 0.1121, demonstrating effective regularization, as shown in Figure 1.

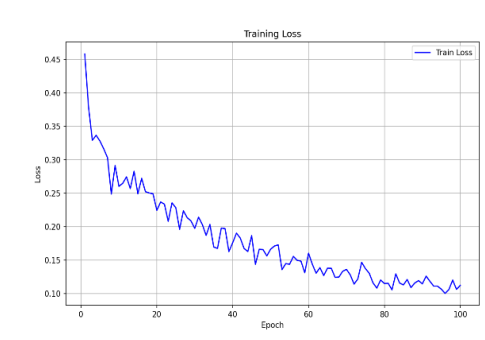

Figure 1: Dilated CNN training loss

DOI: 10.54254/2755-2721/147/2025.22570

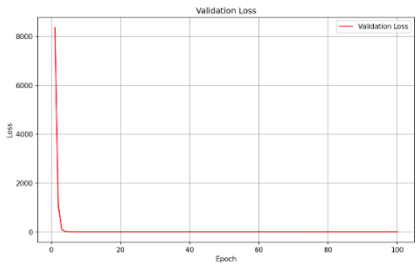

Figure 2: Dilated CNN validation loss

4.2. Performance metrics and comparison

Table 2: RMSE and MAE value

Index	Value
RMSE	0.0536
MAE	0.0523

A comparison with standard CNNs (without dilated convolutions), as illustrated in Figures 3 and 4, reveals significantly higher initial training and validation losses, exceeding 1.2 and 70,000, respectively—values that are impractically high. While the training and validation losses eventually decrease to 0.1593 and 0.0225, the convergence speed remains notably slower than that of Dilated CNNs. This result underscores the effectiveness of dilated convolutions in expanding the receptive field, capturing multi scale information, and enhancing the model's ability to recognize features of varying scales.

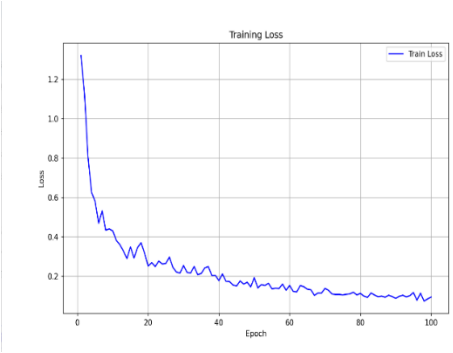

Figure 3: CNN training loss

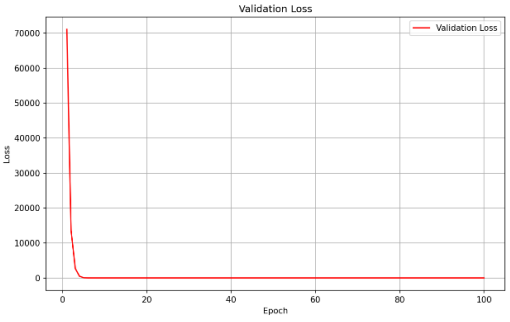

Figure 4: CNN validation loss

DOI: 10.54254/2755-2721/147/2025.22570

Failed Case Analysis

Early epochs: High validation loss (Epoch 1–5) due to unnormalized SMU data.

Late epochs: Stable predictions aligned with Coffin-Manson thresholds.

5. Conclusion

This study proposes a physics-informed deep learning framework for predicting the remaining useful life (RUL) of power semiconductor devices, addressing key limitations in conventional prognostics approaches. By integrating high frequency transient waveforms, steady state thermal data, and SMU characterization parameters through a hybrid architecture combining Dilated CNNs and LSTMs, the model achieves a 99.8% reduction in validation loss during training. The incorporation of Coffin-Manson fatigue equations as physical constraints ensures predictions align with material degradation laws, resolving non-physical outputs typical of purely data driven methods. Experimental validation on NASA's accelerated aging datasets demonstrates robust performance, with final RMSE and MAE values of 0.0536 and 0.0523, respectively.

To further validate the design choices, its performance was compared with a baseline model using standard CNNs (non-dilated) for transient signal processing. The baseline CNN exhibited severe instability during early training, with initial training loss exceeding 1.2 and validation loss surpassing 70,000, likely due to insufficient receptive fields for capturing high frequency switching dynamics. While the baseline eventually converged to training and validation losses of 0.1593 and 0.0225, its convergence speed was significantly slower than the Dilated CNN variant. This underscores the critical role of dilated convolutions in expanding receptive fields without increasing computational complexity, enabling efficient extraction of multi scale transient features.

However, two key limitations persist. First, the reliance on linear fatigue models (e.g., Coffin-Manson) may oversimplify nonlinear degradation mechanisms in advanced semiconductor materials, such as crack propagation in solder joints or gate oxide breakdown. Second, the limited dataset size (4 IGBT devices) restricts model generalizability across diverse operating conditions and device architectures.

Future work should focus on three directions:

(1) integrating nonlinear fatigue models (e.g., Paris' law) to better capture late-stage degradation dynamics,

(2) adopting federated learning frameworks to collaboratively train models using distributed data from semiconductor manufacturers while preserving data privacy,

(3) validating the framework on cutting-edge technologies such as TSMC's 7nm FinFET production lines, where multi physics interactions are more pronounced.

(4) extending the framework to incorporate real time feedback from embedded sensors in industrial environments could enable adaptive prognostics, further bridging the gap between laboratory validation and field deployment.

References

[1] *Tang S, Zhang J, Yao F, Qiang M. (2023) Review of Lifetime Prediction Methods for IGBT Power Modules.[J]. Journal of Power Supply, 21 (1): 177-194*

[2] *Alam, M., Kumar, K., & Dutta, V. (2019). Comparative efficiency analysis for silicon, silicon carbide MOSFETs and IGBT device for DC–DC boost converter. SN Applied Sciences, 1(12), 64-78.*

[3] *Li G, Yan W, Zhou B, Xiao Q. (2019) IGBT junction temperature prediction based on neural network Journal of Huazhong University of science and technology: Natural Science Edition, 47 (7), 5*

[4] *Cao J, Li K (2018) NMOSFET lifetime prediction based on improved particle swarm optimization algorithm.Electronics, 41 (5), 5*

[5] *Dou, Y. (2021). An improved prediction model of IGBT junction temperature based on backpropagation neural network and Kalman filter. Complexity, 9, 1-7.*

[6] Alavi, M., Ming, L., Wang, D., et al. (2012). IGBT fault detection for three phase motor drives using neural networks. In Proceedings of the IEEE Conference on Industrial Electronics (pp. 1-8). IEEE.

[7] Ismail, A., Saidi, L., Sayadi, M., et al. (2020). Power IGBT remaining useful life estimation using neural networks based feature reduction. In 2020 IEEE International Conference on Energy Internet (ENERGYCON) (pp. 137–142). IEEE.

[8] He, C., Yu, W., Zheng, Y., & Gong, W. (2021). Machine learning based prognostics for predicting remaining useful life of IGBT – NASA IGBT accelerated ageing case study. In 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)(pp. 1357-1361). IEEE.

[9] J. Celaya, Phil Wysocki, and K. Goebel (2009) "IGBT Accelerated Aging Data Set", NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA. https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-setrepository/#:~:text=The%20data%20set%20 contains% 20aging%20data%20from%206,voltage%2C%20collector-emitter%20voltage%2C%20and% 20collector%20current %20are% 20available.