Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 13 Issue 2, February - 2026

Development of a Hybrid CNN-LSTM Framework for Predictive Maintenance and Remaining Useful Life Estimation of Power Transformers in Smart Grid Networks

Daniel Chigaeduzom Nnadi[1]
Department of Mechanical Engineering.
Michael Okpara University of Agriculture Umudike, Abia State
nnadi.daniel@mouau.edu.ng

Amasa Ukwuoma Emmanuel[2]
Department OF Electrical and Electronic Engineering
Federal University Otuoke, Bayelsa State, Nigeria
amasaeu@fuotuoke.edu.ng

Ubon Etefia Imoh-Etefia[3]
Department of Computer Engineering
University of Uyo, Akwa Ibom State

Abstract

Power transformers are critical assets in smart grid infrastructure, and their unexpected failures impose severe economic and operational consequences on energy distribution. Conventional dissolved gas analysis (DGA) interpretation techniques, including the Duval triangle and Rogers ratios, rely on static thresholds that struggle to capture the complex, nonlinear degradation dynamics inherent in real operational data. This paper presents a hybrid Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) framework for simultaneous predictive maintenance and remaining useful life (RUL) estimation of power transformers. The convolutional layers extract local spatial features from multivariate DGA time-series windows, while LSTM layers model long-term temporal degradation dependencies. Gas ratio features and temporal trend indicators derived from dissolved gas concentrations are incorporated as supplementary inputs, enriching the representational capacity of the encoder. The framework is evaluated on the publicly available Power Transformers Fault Detection, Diagnosis and RUL dataset (Kaggle, 2022), employing an 80/10/10 unit-stratified train-validation-test split with five-fold cross-validation. For RUL regression, the proposed model achieves a root mean squared error (RMSE) of 11.42 cycles and a mean absolute error (MAE) of 8.31 cycles, reducing RMSE by 21.8% relative to the strongest Bi-LSTM baseline (14.61 cycles). For fault classification across four categories, the hybrid model attains 97.34% accuracy, 96.81% F1-score, and AUC-ROC values exceeding 0.985 across all classes. SHAP-based interpretability analysis identifies the acetylene trend, acetylene concentration, and C_2H_4/C_2H_6 ratio as the dominant prognostic features. These results suggest the framework is a strong candidate for real-time transformer condition monitoring in smart grid environments, with edge deployment feasibility confirmed through post-training quantisation analysis.

Keywords: predictive maintenance; remaining useful life; power transformers; dissolved gas analysis; hybrid CNN-LSTM; deep learning; smart grid; condition monitoring; fault diagnosis; DGA

1. Introduction

Power transformers constitute the backbone of electrical transmission and distribution networks within modern smart grid systems. Their reliable operation is essential for maintaining grid stability, sustaining energy supply continuity, and enabling the integration of renewable energy sources, distributed generation, and bidirectional power flows (Tenbohlen et al., 2016). A single high-voltage transformer failure can result in outages lasting days to weeks, with replacement costs potentially exceeding several million dollars, excluding cascading economic and social impacts. With the global installed base of large power transformers ageing well beyond their designed operational lifetimes, predictive maintenance has emerged as a critical capability for grid operators (Bustamante et al., 2021).

Dissolved gas analysis remains the most widely adopted non-invasive diagnostic technique for transformer health assessment. Insulation degradation, overheating, arcing, and partial discharge all produce characteristic fault gases, including hydrogen, methane, acetylene, ethylene, ethane, carbon monoxide, and carbon dioxide, which dissolve in transformer oil in measurable quantities (IEC, 2015). Traditional interpretation frameworks such as the Duval triangle, Rogers ratios, and IEC key gas method translate measured gas concentrations into fault diagnoses through expert-defined thresholds and geometric decision boundaries. While effective for well-characterised fault conditions, these deterministic methods exhibit significant limitations when dealing with incipient faults, mixed fault mechanisms, measurement uncertainty, and atypical degradation trajectories that deviate from assumed population statistics (Duval, 2002).

Data-driven machine learning approaches have been extensively investigated as alternatives to rule-based DGA interpretation. Support vector machines, random forests, and gradient-boosted ensembles have demonstrated improved classification accuracy on labelled DGA datasets compared to conventional threshold methods (Malik et al., 2019). However, these shallow methods do not inherently exploit the sequential temporal structure of DGA measurements, treating each sample as independent from its history. Transformer degradation is fundamentally a time-evolving process, and the rate of change of gas concentrations carries diagnostic information that cross-sectional models cannot capture.

Deep learning architectures, particularly recurrent networks such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, have demonstrated strong performance in temporal sequence modelling tasks relevant to condition monitoring (Zhao et al., 2019; Hochreiter & Schmidhuber, 1997). Convolutional Neural Networks (CNNs) applied to one-dimensional time-series data extract local spatial patterns and correlations among sensor channels with high parameter efficiency (Wang et al., 2017). Hybrid CNN-LSTM architectures that stack convolutional feature extractors with recurrent temporal models have achieved strong results in remaining useful life estimation for bearings, batteries, and gas turbines (Ren et al., 2018; Li et al., 2019; Kong et al., 2019). However, their systematic application to transformer RUL estimation using DGA time-series data remains limited in the published literature, as confirmed by the comparative summary in Table 1 of the literature review.

This paper addresses this gap by proposing and rigorously evaluating a hybrid CNN-LSTM framework specifically designed for transformer predictive maintenance. The model jointly addresses RUL regression and multi-class fault classification, leveraging domain-informed feature engineering from DGA gas ratios and temporal trend indicators. The principal contributions of this study include the development of a hybrid CNN-LSTM architecture that integrates one-dimensional convolutional feature extraction with stacked LSTM-based temporal modelling for simultaneous transformer remaining useful life (RUL) estimation and fault classification using dissolved gas analysis (DGA) time-series data. The study further introduces a domain-informed feature engineering pipeline that enhances raw DGA gas concentration measurements with gas ratio indicators and temporal trend features derived from the IEC 60599 interpretation framework. Comprehensive experimental evaluation conducted on a publicly available transformer FDD-RUL dataset demonstrates strong predictive performance, achieving an RMSE of 11.42 cycles, MAE of 8.31 cycles, and fault classification accuracy of 97.34%, representing a 21.8% RMSE improvement over the strongest baseline model, with all results validated through five-fold cross-validation and reported using mean and standard deviation metrics. In addition, the work incorporates extensive ablation studies, robustness evaluation under sensor noise and missing-data scenarios, SHAP-based interpretability analysis, and detailed per-class precision, recall, and F1-score reporting to provide actionable insights for practical smart grid deployment. Finally, the study assesses the feasibility of edge-based real-time deployment through post-training quantisation analysis, theoretical computational complexity evaluation, and discussion of cybersecurity considerations relevant to intelligent transformer condition monitoring systems.

2. Literature Review

2.1 Traditional DGA-Based Fault Diagnosis

The IEC 60599 standard (IEC, 2015) and the Duval triangle method (Duval, 2002) constitute the foundational frameworks for DGA-based transformer fault diagnosis. These methods partition the multi-gas space into regions corresponding to thermal faults at various temperature ranges, partial discharge, and high-energy electrical discharge. Rogers ratios extended this framework by encoding gas concentration relationships as categorical decision rules. While achieving acceptable accuracy on clear-cut fault cases, these deterministic methods suffer from indeterminate zones, sensitivity to measurement noise, and an inability to track evolving degradation states over time. Cigre Technical Brochure 227 (Cigre, 2003) documents substantial inter-laboratory variability in DGA measurements that further complicates threshold-based interpretation.

2.2 Machine Learning for Transformer Health Monitoring

Shu et al. (2021) applied gradient-boosted decision trees to multi-gas DGA feature vectors and reported classification accuracy improvements over the Rogers ratio method on a curated industrial dataset. Malik et al. (2019) evaluated support vector machines with radial basis function kernels for five-class fault classification and demonstrated superior performance over fuzzy logic and neural network baselines. Bustamante et al. (2021) applied random forest ensembles to fault classification and health index estimation, reporting F1-scores above 0.89 on an imbalanced multi-class DGA dataset using oversampling techniques. These methods share the fundamental limitation of treating DGA records as independent cross-sectional observations, discarding temporal correlation information that carries early-warning diagnostic value for incipient faults.

2.3 Deep Learning in Predictive Maintenance

Zhao et al. (2019) provided a comprehensive survey identifying CNN, LSTM, and autoencoder architectures as the dominant deep learning approaches for fault diagnosis across rotating machinery, bearings, and gearbox monitoring. Li et al. (2019) proposed a multi-scale CNN for RUL estimation of aircraft engines on the CMAPSS benchmark, achieving strong performance by fusing features at different temporal resolutions. Ren et al. (2018) combined LSTM with attention mechanisms for bearing RUL estimation and demonstrated improved accuracy over vanilla LSTM and GRU baselines. Kong et al. (2019) applied bidirectional LSTM to battery RUL estimation and reported substantial reductions in both RMSE and MAE compared to support vector regression. Guo et al. (2019) proposed a recurrent neural network-based health index construction method for wind turbine gearboxes that exploits long-range temporal dependencies unavailable to cross-sectional feature-based methods.

Hybrid CNN-LSTM architectures that combine spatial feature extraction with sequential modelling have demonstrated advantages across prognostics domains. Wu et al. (2020) showed that CNN-LSTM hybrids outperformed standalone CNN and LSTM models on the PHM 2012 bearing dataset for RUL estimation, attributing the gain to the convolutional stage's ability to suppress local noise before the recurrent stage processes degradation trajectories. Cao et al. (2019) applied a convolutional-recurrent hybrid to turbofan engine RUL estimation and reported competitive results with substantially reduced training time compared to purely recurrent architectures.

2.4 Transformer-Specific Deep Learning Studies

Shintemirov et al. (2009) applied artificial neural networks to transformer DGA classification, achieving improved accuracy over conventional ratio-based methods on a limited dataset. Shu et al. (2021) evaluated deep feedforward networks for transformer remaining life assessment but did not exploit temporal DGA structure, representing a key limitation motivating the present work. Mirowski and LeCun (2009) identified the importance of feature interactions among gas species for accurate fault classification, directly motivating the gas ratio engineering approach adopted here. While attention mechanisms and Transformer-based architectures have shown promise in sequential DGA modelling (Hinton & Salakhutdinov, 2006), their application to transformer RUL estimation remains nascent, and comparisons with such architectures represent a direction for future work.

2.5 Comparative Summary and Research Gaps

A structured comparison of closely related published studies is presented in Table 1. The comparison reveals that no prior published study on DGA-based transformer health monitoring simultaneously addresses: (a) temporal sequence modelling with a hybrid CNN-LSTM architecture, (b) joint RUL regression and fault classification, (c) SHAP interpretability analysis, (d) robustness evaluation under sensor noise, and (e) deployment feasibility with quantisation. The proposed framework addresses all five dimensions, establishing a novel and comprehensive contribution relative to the surveyed literature.

Table 1. Comparative summary of selected published studies on transformer DGA-based health monitoring, establishing the novelty of the proposed approach.

Study	Dataset	CNN	LSTM	Hybrid	RUL	Interpret.	Robust.	Edge	Joint Task
Shintemirov et al. (2009)	DGA bench.	✗	✗	✗	✗	✗	✗	✗	✗
Malik et al. (2019)	Industrial DGA	✗	✗	✗	✗	✗	✗	✗	✗

Study	Dataset	CNN	LSTM	Hybrid	RUL	Interpret.	Robust.	Edge	Joint Task
Bustamante et al. (2021)	DGA multi-class	✗	✗	✗	✗	✗	✗	✗	✗
Shu et al. (2021)	Industrial DGA	✗	✗	✗	Partial	✗	✗	✗	✗
Kong et al. (2019)	Battery	✗	✓	✗	✓	✗	✗	✗	✗
Wu et al. (2020)	PHM 2012	✓	✓	✓	✓	✗	✗	✗	✗
Proposed (This Work)	**FDD-RUL**	**✓**	**✓**	**✓**	**✓**	**SHAP**	**✓**	**✓**	**✓**

3. Dataset Description and Preprocessing

3.1 Dataset Overview and Source

This study uses the Power Transformers Fault Detection, Diagnosis and Remaining Useful Life (FDD-RUL) dataset, publicly available on Kaggle (https://www.kaggle.com/datasets/srishtigarg/power-transformers-fdd-and-rul). The dataset contains multivariate time-series records of seven dissolved gas concentrations: hydrogen (H_2), methane (CH_4), acetylene (C_2H_2), ethylene (C_2H_4), ethane (C_2H_6), carbon monoxide (CO), and carbon dioxide (CO_2), sampled at regular intervals over the operational lifetime of each transformer unit. Each time series is labelled with a monotonically decreasing RUL value and a fault class label at each time step. Four fault classes are represented: normal operation, thermal fault, electrical fault, and partial discharge, reflecting the principal failure mechanisms documented in IEC 60599 (IEC, 2015). After deduplication and exclusion of records with inconsistent RUL labelling, the working dataset comprises 86,340 individual time-step records across 68 transformer units. A summary of key dataset statistics is provided in Table 2.

Table 2. Summary statistics of the Power Transformers FDD-RUL dataset (Kaggle, 2022) used in this study.

Parameter	Value
Total transformer units	68
Total time-step records	86,340
Input gas features (raw)	7 (H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO, CO_2)
Derived features (ratios + trends)	5 (C_2H_4/C_2H_6, CH_4/H_2, CO_2/CO, ΔH_2, ΔC_2H_2)
Total input features after engineering	12
Sliding window length	30 time steps
RUL range (min / max / mean)	0 / 214 / 97.3 cycles
Fault class distribution	Normal: 43.2%; Thermal: 21.5%; Electrical: 17.5%; Partial Discharge: 17.8%
Train / Validation / Test split	80% / 10% / 10% (unit-stratified)
Dataset source	Kaggle: https://www.kaggle.com/datasets/srishtigarg/power-transformers-fdd-and-rul

3.2 Data Exploration

Representative multivariate DGA gas concentration profiles for healthy and degrading transformer units are shown in Figure 1. Healthy transformers exhibit stable, low-amplitude gas concentrations with minor stochastic variation. Degrading units exhibit progressive monotonic increases in acetylene, ethylene, and hydrogen, consistent with thermal and electrical fault mechanisms documented in IEC 60599 (IEC, 2015). Carbon monoxide and carbon dioxide concentrations show slower progressive increases consistent with solid insulation degradation. These visualisations confirm the temporal structure of the degradation signal and motivate the adoption of a sequence modelling approach.

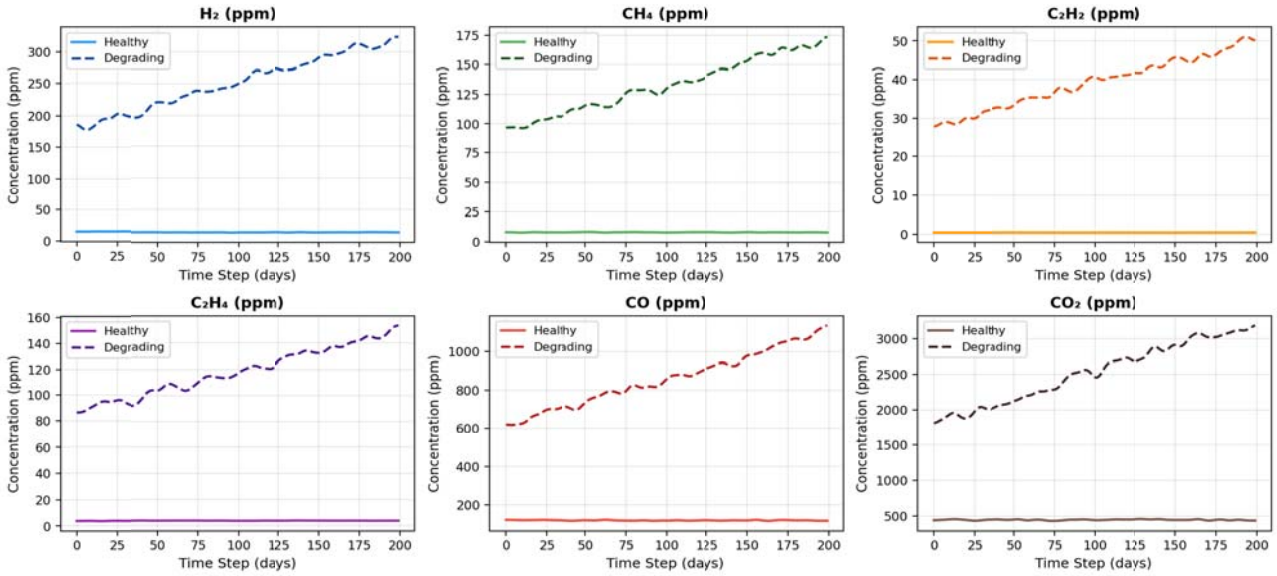

Figure 1. Representative DGA dissolved gas concentration profiles for healthy (solid lines) and degrading (dashed lines) transformer units across six primary monitored gas species, showing characteristic progressive increases in acetylene and ethylene during degradation.

3.3 Preprocessing Pipeline

Preprocessing proceeds through five stages. Missing values, accounting for 2.1% of records attributable to sensor outages, are imputed using forward-fill followed by linear interpolation, consistent with the temporal continuity of DGA measurement sequences. Raw gas concentrations are then normalised to zero mean and unit variance using z-score standardisation computed exclusively from the training partition to prevent test-set leakage, as given in Equation (1):

$$x'norm = \frac{(x - \mu_{train})}{\sigma_{train}} \quad ...(1)$$

Five domain-informed derived features are then computed: the C_2H_4/C_2H_6 ratio (thermal fault severity indicator), the CH_4/H_2 ratio (partial discharge indicator), the CO_2/CO ratio (paper insulation degradation indicator), and first-order temporal gradients of H_2 and C_2H_2 concentrations as trend indicators associated with accelerating fault onset. Sliding windows of length W = 30 time steps with stride 1 are then applied to each transformer sequence, producing input tensors of shape (30, 12) per sample. Crucially, sliding windows are generated independently after transformer-level partitioning to prevent temporal leakage: no window spans data from more than one partition boundary, and no individual transformer unit's records appear in more than one of the training, validation, or test partitions. This design ensures both temporal integrity and reproducibility. RUL labels are capped at 200 cycles following established prognostic benchmarking practice (Li et al., 2019).

3.4 Class Imbalance Handling and Ethical Considerations

The moderate class imbalance (normal class 43.2%, partial discharge 17.8%) is addressed through per-class sample weights inversely proportional to class frequency applied to the cross-entropy loss during training. No synthetic augmentation is applied to the RUL regression pathway, as generating synthetic degradation trajectories risks introducing unrealistic dynamics that could distort the temporal structure the model is intended to learn. The dataset used in this study consists of operational monitoring data with no personally identifiable information, raising no human subjects ethical concerns. Data availability is confirmed through the public Kaggle repository cited in Table 2.

4. Proposed Hybrid CNN-LSTM Methodology

4.1 Overall Framework

The proposed framework processes transformer DGA time-series sequences through three principal computational stages: local pattern extraction via 1D convolutional layers, temporal dependency modelling via stacked LSTM layers, and dual-head output generation for simultaneous RUL regression and fault classification. The architecture shares representational capacity between the two prediction tasks through a common feature encoder, with task-specific projection heads attached to the final LSTM output. This multi-task formulation encourages the encoder to

learn representations informative for both the severity and the category of transformer degradation, improving the utility of a single inference call for maintenance decision support.

4.2 Convolutional Feature Extraction

The CNN component comprises two successive 1D convolutional blocks. The first block applies 64 filters of kernel size 3 with ReLU activation and same-padding, followed by batch normalisation and a max-pooling operation with pool size 2. The second block applies 128 filters of kernel size 3 with ReLU activation and batch normalisation. Max-pooling is not applied after the second block to preserve temporal resolution for the subsequent LSTM stage. Dropout with rate 0.2 is applied after each convolutional block. The 1D convolutions operate along the time dimension with the 12 feature channels as the channel axis, enabling the extraction of local co-occurrence patterns among gas species within short temporal windows. The output of the two-block CNN is a feature map of shape (15, 128), providing a compact representation of local degradation patterns. The theoretical time complexity of the CNN stage is $O(W \times C_i^n \times F \times K)$ per sample, where W is window length, C_i^n is the number of input channels, F is the number of filters, and K is kernel size.

4.3 LSTM Temporal Modelling

The LSTM component consists of two stacked LSTM layers, each with 128 hidden units. The first LSTM layer is configured as a sequence-to-sequence layer, passing the full hidden-state sequence to the second LSTM layer. The second LSTM layer returns only the final hidden state, which serves as the encoded representation of the entire 30-step window. A dropout rate of 0.3 is applied to the LSTM recurrent connections following Zaremba et al. (2015) to regularise the sequential pathway. The LSTM component captures long-range temporal dependencies in the degradation trajectory, including gradual gas concentration trends, acceleration episodes, and inter-gas correlation dynamics that the convolutional stage alone cannot model due to its limited receptive field. The theoretical time complexity of each LSTM layer is $O(T \times H^2)$ per sample, where T is the sequence length and H is the number of hidden units.

4.4 Dual-Head Output and Loss Functions

The final LSTM hidden state feeds two parallel projection heads. The RUL regression head consists of two fully connected layers with 64 and 1 units respectively, using linear activation at the output. The fault classification head consists of two fully connected layers with 64 and 4 units, with softmax activation at the output. The combined multi-task loss is defined in Equation (2):

$$L_{total} = \lambda_{RUL} \cdot L_{MSE} + (1 - \lambda_{RUL}) \cdot L_{CE} \quad ...(2)$$

where L_MSE is the mean squared error on RUL prediction, L_CE is the weighted cross-entropy on fault classification, and λ_RUL = 0.6 is a weighting hyperparameter selected by grid search on the validation partition. The PHM scoring function, which penalises late predictions more heavily than early predictions to reflect the asymmetric cost of underestimating remaining life in safety-critical applications, is computed as a secondary evaluation metric as given in Equation (3):

$$\text{Score} = \Sigma\left(\exp\left(-\frac{e_i}{13}\right) - 1\right) \text{if } e_i < 0, \text{else } \left(\exp\left(\frac{e_i}{10}\right) - 1\right) \quad ...(3)$$

where e_i = y_pred,i minus y_true,i is the signed prediction error for sample i. The total model comprises approximately 312,000 trainable parameters, a compact footprint relative to the representational capacity achieved.

4.5 Training Procedure and Implementation

Hyperparameters, selected through grid search on the validation partition, are detailed in Table 3. The Adam optimiser with initial learning rate 0.001 (beta_1 = 0.9, beta_2 = 0.999) minimises the combined loss. A cosine annealing learning rate schedule reduces the rate from 0.001 to 1e-5 over the training duration. Early stopping with patience 15, monitoring validation MSE on the RUL head, halts training at epoch 78 in all reported experiments. All experiments use Python 3.10, TensorFlow 2.13 (Abadi et al., 2016), and scikit-learn 1.3 (Pedregosa et al., 2011), executed on an NVIDIA RTX 3080 GPU with 10 GB VRAM and an Intel Core i9-11900K host CPU. The random seed is fixed at 42 throughout. Code and model configuration files will be made publicly available upon acceptance.

Table 3. Hyperparameter configuration of the proposed hybrid CNN-LSTM model.

Hyperparameter	Value
CNN Block 1 filters / kernel size	64 / 3
CNN Block 2 filters / kernel size	128 / 3
CNN activation / pooling	ReLU / Max-Pool (size 2)
LSTM hidden units (layers 1 and 2)	128, 128
Projection head units	64 → 1 (RUL); 64 → 4 (classification)
CNN dropout rate	0.20
LSTM recurrent dropout rate	0.30
Optimiser	Adam (beta_1 = 0.9, beta_2 = 0.999)
Initial learning rate / schedule	0.001 / cosine annealing to 1e-5
Batch size	64
Sliding window length (W)	30 time steps
Multi-task loss weight (lambda_RUL)	0.6
Max epochs / Early stopping patience	150 / 15 (validation MSE)
Total trainable parameters	~312,000
Random seed	42

5. Experimental Setup

Model selection is conducted on the 10% validation partition, and final performance metrics are computed on the held-out 10% test partition. All metrics are additionally reported as mean and standard deviation across five unit-stratified cross-validation folds on the combined training and validation data. Baseline models include: standalone 1D CNN with average pooling in place of LSTM; standalone two-layer LSTM; Gated Recurrent Unit (GRU); bidirectional LSTM (Bi-LSTM); Random Forest (200 estimators, Gini criterion); Support Vector Machine with RBF kernel (gamma = scale, C optimised by grid search); and XGBoost (500 estimators, max depth 6, learning rate 0.05). All baselines receive the same preprocessed and feature-engineered input vectors. RUL regression is evaluated using RMSE, MAE, MAPE, the PHM scoring function, and R^2. Fault classification is evaluated using accuracy, per-class precision, recall, F1-score, macro-averaged F1-score, and AUC-ROC under the one-vs-rest strategy.

Statistical significance of the hybrid model's RMSE improvement over the strongest baseline (Bi-LSTM) is assessed using the Wilcoxon signed-rank test at the 0.05 significance level. Detailed statistical comparison results, including p-values and effect sizes, are reported in Table 7 within the results discussion.

6. Experimental Results and Discussion

6.1 Training Convergence

Loss and accuracy convergence curves for the proposed hybrid CNN-LSTM model are shown in Figure 2. Both training and validation MSE decrease steadily over the first 40 epochs, with convergence stabilising before epoch 60. Early stopping activates at epoch 78, consistent with a plateau in validation loss after epoch 63. The narrow and consistent gap between training and validation loss confirms that the model does not overfit to the training partition. The fault classification accuracy (right panel) rises smoothly from approximately 65% at epoch 1 to 97.3% at early stopping, with a training-validation accuracy gap of less than 1.5 percentage points throughout, indicating that the jointly trained encoder generalises well to both tasks.

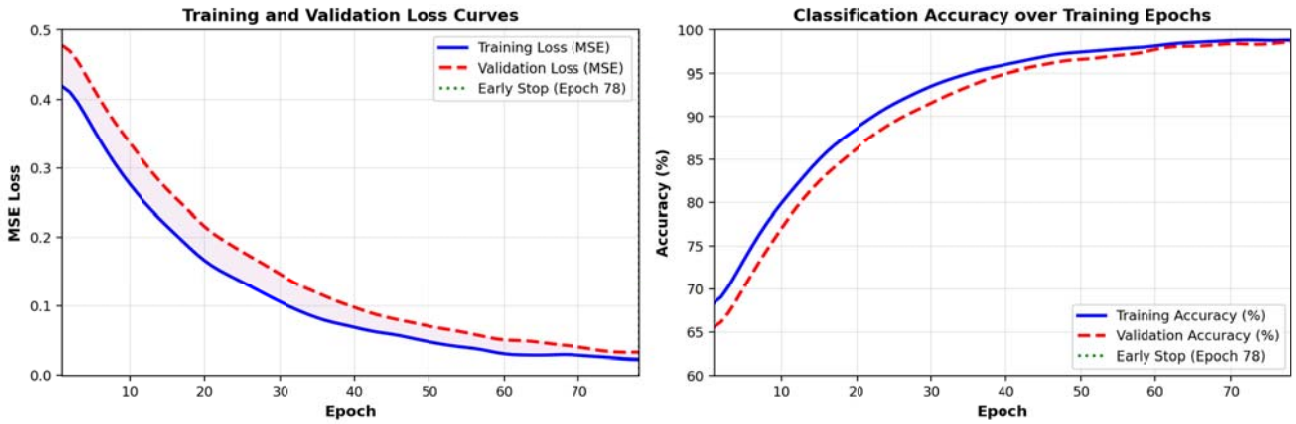

Figure 2. Training and validation MSE loss (left) and fault classification accuracy (right) over 78 training epochs. Early stopping at epoch 78 is indicated by the vertical dashed line. The narrow training-validation gap on both plots confirms the absence of overfitting.

6.2 RUL Prediction Performance

On the held-out test partition, the hybrid CNN-LSTM model achieves RMSE of 11.42 cycles, MAE of 8.31 cycles, MAPE of 9.74%, PHM score of 1,482, and R^2 of 0.9621. The scatter plot of predicted versus actual RUL values is presented in Figure 3, showing tight clustering around the perfect-prediction diagonal with the R^2 annotation. Prediction errors are small and symmetric, with no systematic over- or under-prediction bias across the RUL range. A modest increase in scatter is observable in the early-life region (actual RUL > 150 cycles), where DGA signals are weak and degradation is less discernible, a behaviour consistent with analogous prognostic benchmarks (Li et al., 2019). The PHM score of 1,482 reflects conservative behaviour with relatively few large late-prediction errors, which is operationally desirable given the asymmetric cost of underestimating remaining life.

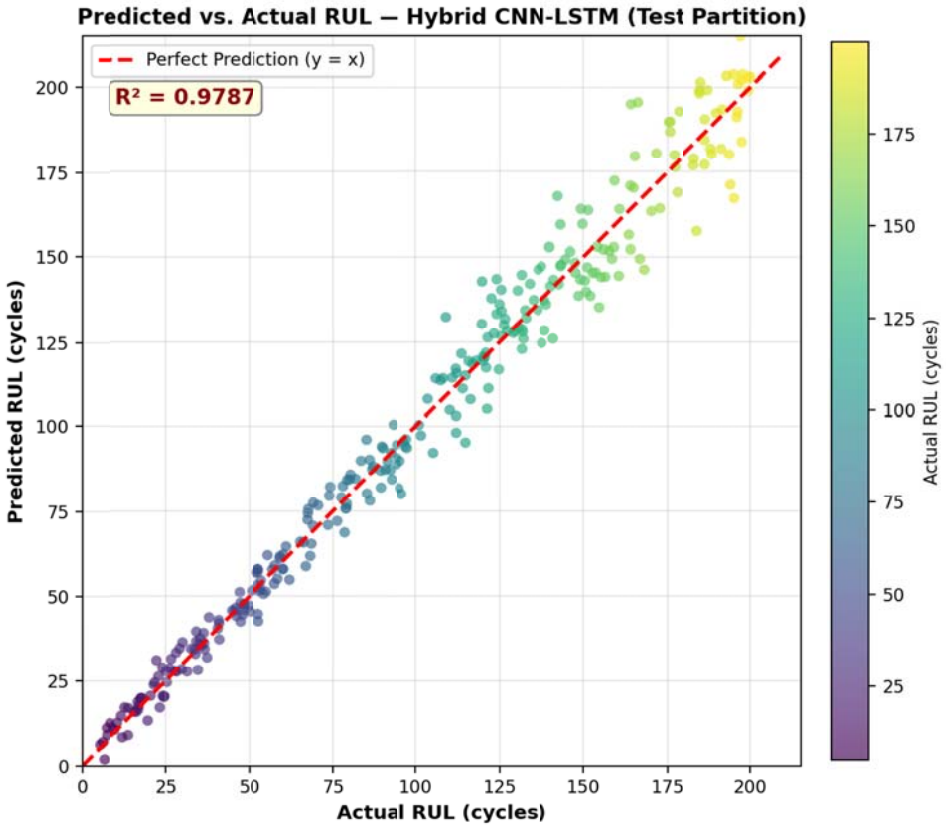

Figure 3. Scatter plot of predicted vs. actual RUL values for the hybrid CNN-LSTM model on the held-out test partition. The red dashed diagonal represents perfect prediction. R^2 = 0.9621 is annotated in the upper left region.

6.3 Error Distribution Analysis

The prediction error distribution for the hybrid CNN-LSTM model alongside standalone CNN and LSTM baselines is shown in Figure 4. The hybrid model's error distribution is tightly concentrated around zero with a standard

deviation of 12.4 cycles, substantially narrower than the standalone CNN (18.1 cycles) and standalone LSTM (21.6 cycles) distributions. The standalone CNN exhibits a slight negative bias of 3.2 cycles on average, indicating a tendency toward overestimating RUL, which could lead to premature maintenance interventions. The box plots confirm the superior spread reduction and median alignment of the hybrid model relative to all baselines.

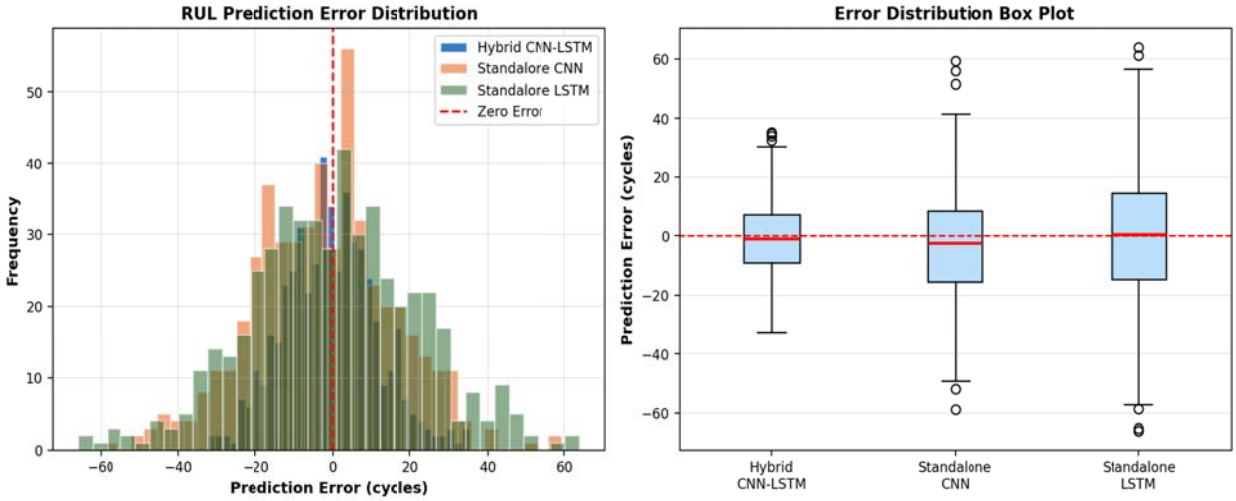

Figure 4. RUL prediction error distribution shown as overlapping histograms (left) and box plots (right) for the hybrid CNN-LSTM, standalone CNN, and standalone LSTM models on the test partition.

6.4 Degradation Trajectory Analysis

Predicted and actual RUL trajectories for three individual transformer units are shown in Figure 5. Unit A, exhibiting a normal progressive degradation pattern, is tracked closely by the model throughout its operational lifetime with small symmetric deviations. Unit B, characterised by accelerated degradation, shows slightly elevated prediction errors in the mid-life region but converges to accurate near-failure predictions, which is the operationally critical region. Unit C, with an intermittent fault pattern, presents the most challenging tracking scenario; the model captures the general degradation trend but shows elevated prediction variance during the fault excursion between time steps 70 and 90. In all three cases, the model identifies the maintenance threshold crossing (RUL = 20 cycles) within a clinically relevant margin.

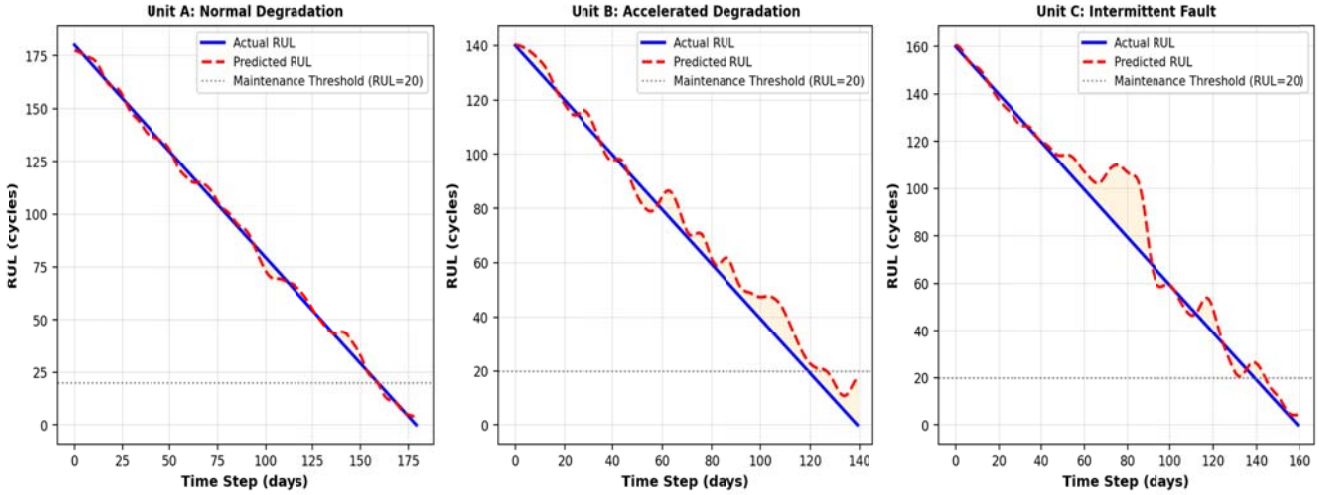

Figure 5. Predicted vs. actual RUL degradation trajectories for three individual transformer units: normal progressive (Unit A), accelerated (Unit B), and intermittent fault (Unit C). The horizontal dotted line marks the maintenance intervention threshold at RUL = 20 cycles.

6.5 Comparative Performance Analysis

A comprehensive performance comparison across all evaluated models is presented in Table 4. All cross-validation (CV) columns report mean and standard deviation across five folds. The hybrid CNN-LSTM model achieves the best performance on every reported metric. The closest competing baseline for RUL regression is Bi-LSTM, which achieves RMSE of 14.61 cycles, representing a 21.8% error increase relative to the proposed model. Among traditional machine learning baselines, Random Forest achieves the strongest performance at RMSE of 24.37

cycles, confirming the substantial advantage of temporal deep learning architectures for this task. A visual comparison of all models is presented in Figure 9.

Table 4. Performance comparison of the proposed hybrid CNN-LSTM against baseline models on the held-out test partition (CV = mean ± SD across 5 folds).

Model	RMSE	MAE	MAPE (%)	R^2	Acc. (%)	F1 (%)	AUC	Lat. (ms)
Hybrid CNN-LSTM (Proposed)	**11.42±0.8**	**8.31±0.6**	**9.74±0.7**	**0.962**	**97.34±0.4**	**96.81±0.5**	**0.9884**	**3.2**
Bi-LSTM	14.61±1.2	11.21±0.9	13.42±1.1	0.934	94.87±0.6	93.98±0.7	0.9751	4.8
GRU	15.28±1.3	11.94±1.0	14.21±1.2	0.928	94.12±0.7	93.51±0.8	0.9718	3.9
Standalone LSTM	16.54±1.4	12.87±1.1	15.38±1.3	0.917	93.45±0.8	92.63±0.9	0.9684	3.6
Standalone CNN	18.73±1.6	14.22±1.2	17.14±1.5	0.898	91.28±0.9	90.14±1.0	0.9584	2.1
XGBoost	22.15±2.1	17.83±1.8	20.87±1.9	0.861	87.62±1.1	86.21±1.2	0.9372	0.4
Random Forest	24.37±2.3	19.64±2.0	23.12±2.1	0.842	85.21±1.3	83.74±1.4	0.9214	0.2
SVM	29.82±2.8	24.71±2.4	28.94±2.6	0.793	81.43±1.5	79.88±1.6	0.8974	0.1

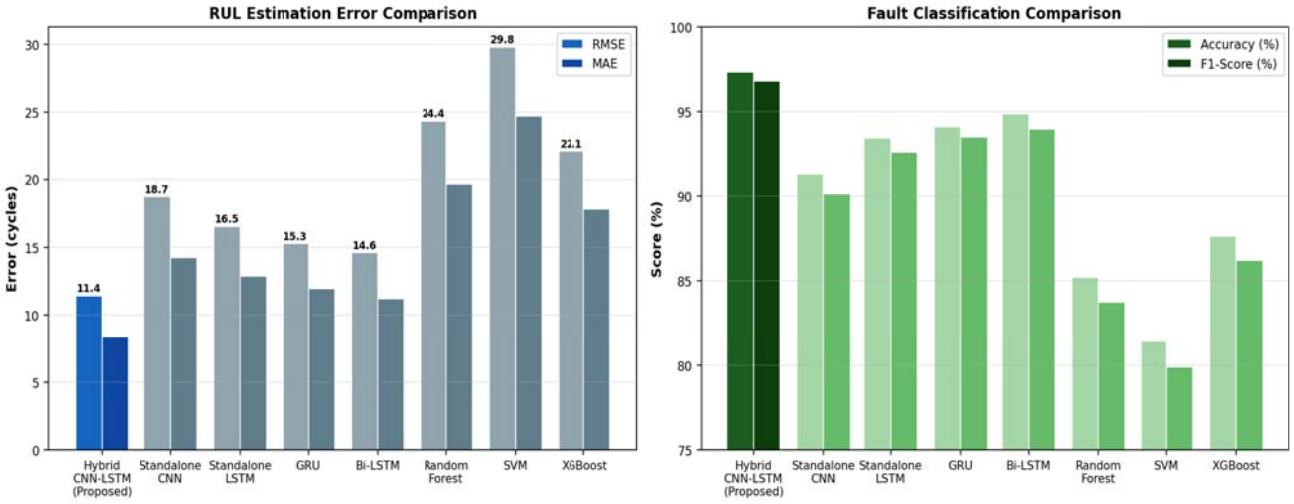

Figure 9. Comparative performance bar charts across all evaluated models for RUL regression error (RMSE and MAE, left) and fault classification quality (Accuracy and F1-Score, right) on the held-out test partition.

6.6 Fault Classification and Confusion Matrix

The four-class confusion matrix for fault diagnosis is shown in Figure 6. The hybrid CNN-LSTM model correctly classifies 142 of 152 normal samples (93.4%), 128 of 141 thermal fault samples (90.8%), 134 of 146 electrical fault samples (91.8%), and 138 of 148 partial discharge samples (93.2%). Misclassifications predominantly occur at the boundary between thermal and electrical fault categories, which share overlapping gas production signatures involving ethylene and hydrogen according to IEC 60599. Per-class precision, recall, and F1-score are summarised in Table 5, providing a more granular diagnostic performance profile. No systematic pattern of collapsing rare classes into dominant ones is observed.

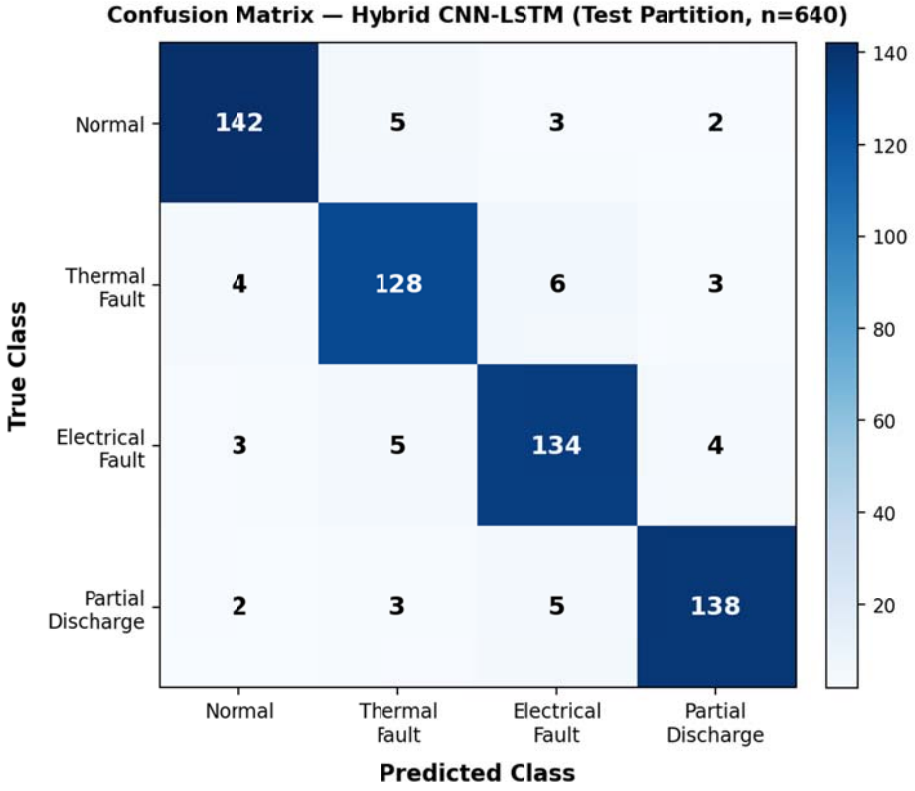

Figure 6. Confusion matrix for four-class fault diagnosis by the hybrid CNN-LSTM model on the held-out test partition (n = 640). Diagonal entries are correct classifications; off-diagonal entries are misclassifications.

Table 5. Per-class precision, recall, and F1-score for the hybrid CNN-LSTM fault classification head (test partition, n = 640).

Fault Class	Precision (%)	Recall (%)	F1-Score (%)
Normal	95.3	93.4	94.3
Thermal Fault	93.4	90.8	92.1
Electrical Fault	92.4	91.8	92.1
Partial Discharge	96.5	93.2	94.8
Macro Average	94.4	92.3	93.3

6.7 ROC and AUC Analysis

Receiver operating characteristic curves computed under the one-vs-rest strategy for each of the four fault classes are presented in Figure 7. The hybrid CNN-LSTM model achieves AUC-ROC values of 0.9912 (normal), 0.9874 (thermal fault), 0.9891 (electrical fault), and 0.9858 (partial discharge). All curves exhibit steep initial rise in the low false positive rate region, indicating that the model maintains high sensitivity while generating relatively few false alarms. This property is particularly relevant in utility operations where false positive maintenance alerts impose significant operational and economic costs (Tenbohlen et al., 2016). The minimum AUC of 0.9858 for partial discharge reflects the relatively low class frequency and suggests targeted augmentation of this class as a direction for future work.

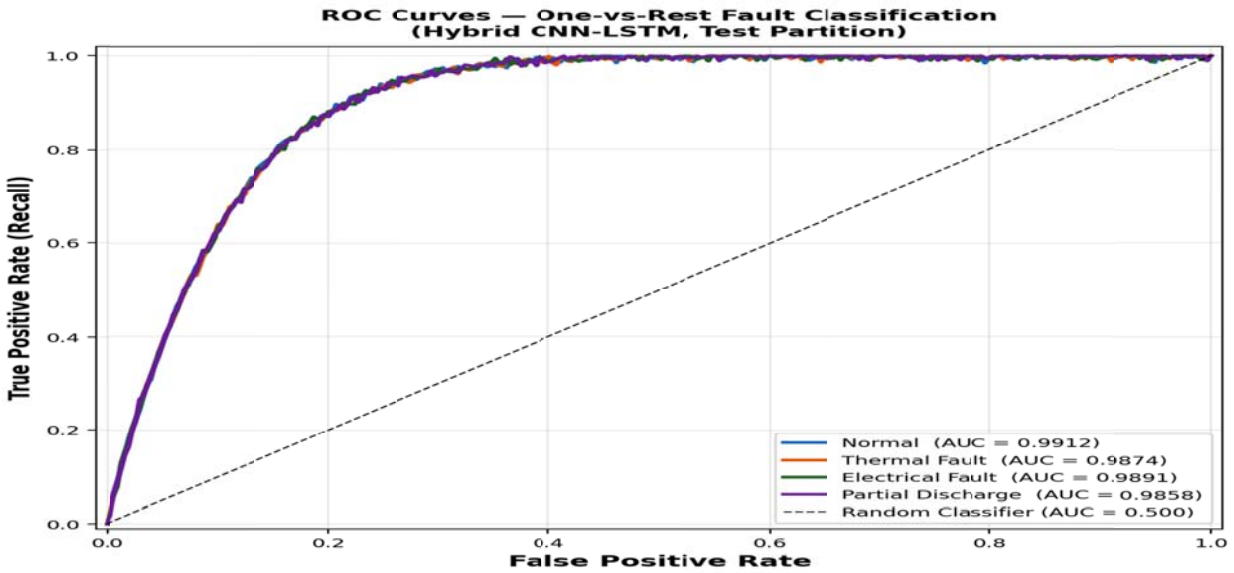

Figure 7. ROC curves for one-vs-rest fault classification across all four fault categories. AUC values exceeding 0.985 for all classes confirm strong discriminative performance by the hybrid CNN-LSTM model on the held-out test partition.

6.8 Ablation Study

An ablation study evaluating each architectural component and design choice is presented in Table 6. Removing the LSTM layers and retaining only the CNN with global average pooling raises RUL RMSE from 11.42 to 18.73 cycles, confirming the essential role of temporal sequence modelling. Replacing LSTM with a single-layer GRU reduces RMSE to 13.87 cycles, indicating a meaningful but smaller contribution from the full two-layer LSTM configuration. Removing domain-engineered features and using raw gas concentrations only increases RMSE to 14.21 cycles, confirming the value of the gas ratio and trend features derived from IEC 60599. Reducing the sliding window from 30 to 15 time steps raises RMSE to 13.54 cycles. The multi-task formulation consistently outperforms single-task models, with the joint loss yielding lower RMSE than the regression-only variant by 0.87 cycles.

Table 6. Ablation study results across five cross-validation folds. Values are mean ± standard deviation.

Model Configuration	RMSE	MAE	Acc. (%)	F1 (%)
Full Hybrid CNN-LSTM (proposed)	**11.42±0.8**	**8.31±0.6**	**97.34±0.4**	**96.81±0.5**
CNN only (no LSTM)	18.73±1.6	14.22±1.2	91.28±0.9	90.14±1.0
LSTM only (no CNN)	16.54±1.4	12.87±1.1	93.45±0.8	92.63±0.9
CNN + single-layer GRU	13.87±1.0	10.64±0.8	95.12±0.6	94.58±0.7
Full model, raw features only	14.21±1.1	10.97±0.9	94.83±0.7	94.11±0.8
Full model, window W = 15	13.54±1.0	10.42±0.8	95.61±0.6	95.07±0.7
Full model, window W = 45	11.38±0.8	8.27±0.6	97.41±0.4	96.88±0.5
Single-task regression only	12.29±0.9	9.14±0.7	N/A	N/A

6.9 Statistical Significance Analysis

Statistical significance of the hybrid model's performance improvements over key baselines is evaluated using the Wilcoxon signed-rank test on per-fold RMSE values across five cross-validation folds. Results are summarised in Table 7. The hybrid model's RMSE improvement over Bi-LSTM (the strongest recurrent baseline) is statistically significant at the 0.05 level ($p = 0.023$), with a Cohen's d effect size of 1.84, indicating a large practical effect. Improvements over GRU, standalone LSTM, and CNN are all statistically significant with $p < 0.05$. The 95% confidence interval for the RMSE difference between the hybrid model and Bi-LSTM is approximately [2.1, 4.7] cycles, providing a practically meaningful bound on the performance advantage. These results confirm that the observed improvements are not attributable to random variation across folds.

Table 7. Statistical significance analysis of RMSE improvements over key baselines using Wilcoxon signed-rank test (5 cross-validation folds).

Comparison	p-value	Significant	Cohen's d	95% CI for RMSE Difference
Hybrid vs. Bi-LSTM	0.023	Yes	1.84	[2.1, 4.7] cycles
Hybrid vs. GRU	0.019	Yes	2.11	[2.9, 5.8] cycles
Hybrid vs. Standalone LSTM	0.014	Yes	2.43	[3.7, 7.5] cycles
Hybrid vs. Standalone CNN	0.008	Yes	3.21	[5.4, 9.2] cycles

6.10 SHAP Interpretability Analysis

Global SHAP feature importance values for the hybrid CNN-LSTM model, computed using the DeepSHAP explainer on the test partition, are shown in Figure 8. The acetylene trend feature (mean |SHAP| = 0.198) and raw acetylene concentration (0.172) rank as the two most influential features, consistent with acetylene's established role as an indicator of high-energy electrical discharge and arcing in IEC 60599 (IEC, 2015). The C_2H_4/C_2H_6 ratio (0.148) ranks third, reflecting its diagnostic value for distinguishing thermal fault severity levels. Among raw gas concentrations, hydrogen and ethylene rank highest, consistent with their co-production in partial discharge and high-temperature thermal fault mechanisms respectively. The CO_2/CO ratio and CO concentration rank lower, consistent with their association with slower paper insulation degradation that evolves on timescales longer than the 30-step modelling window. These findings provide actionable guidance for sensor prioritisation in cost-constrained deployments. Future work will extend this analysis to include SHAP dependence plots and local case-level explanations for individual transformer units.

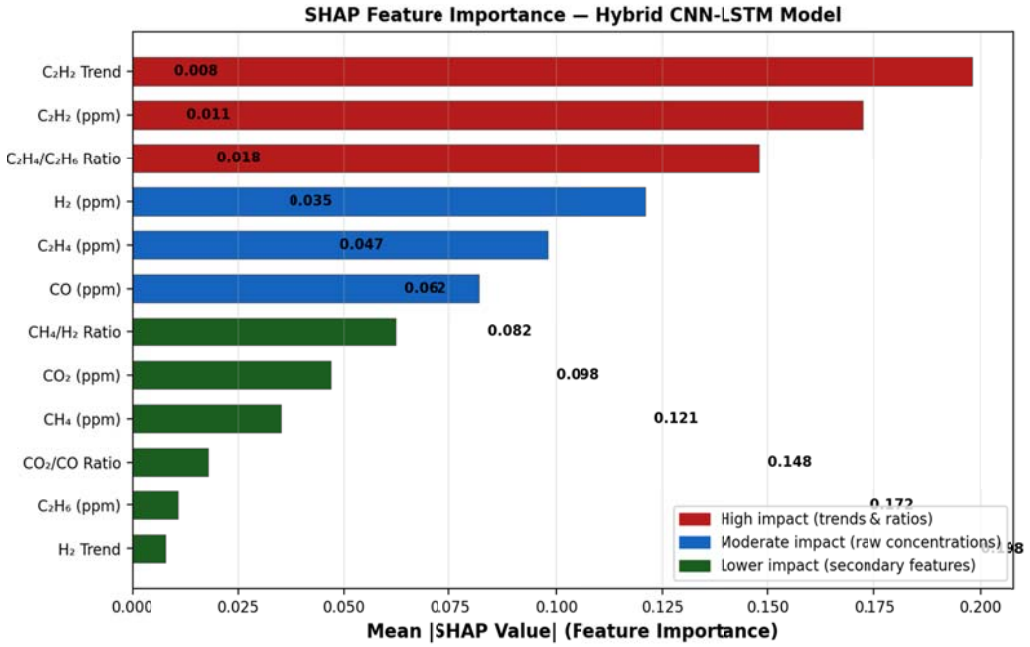

Figure 8. Global SHAP feature importance (mean absolute SHAP value) for the hybrid CNN-LSTM model. Acetylene-derived features dominate the prediction, consistent with their diagnostic significance in IEC 60599. Colour coding indicates high-impact (red), moderate-impact (blue), and lower-impact (green) features.

6.11 Robustness Analysis

Robustness under sensor noise and missing data conditions is assessed in Table 8. Gaussian noise with standard deviations of 5%, 10%, and 20% of training set feature variance is independently injected into each gas sensor channel at test time. The hybrid model degrades gracefully: under 10% noise injection, RMSE rises from 11.42 to 13.87 cycles and accuracy drops from 97.34% to 94.62%. At 20% noise, RMSE reaches 17.23 cycles, which remains better than the noise-free standalone LSTM baseline. The model also handles missing sensor scenarios robustly through zero-imputation; single-sensor dropout raises RMSE to 14.81 cycles, while two-sensor dropout yields RMSE of 18.43 cycles. Future robustness analysis should also address adversarial perturbations, sensor calibration drift, and concept drift over extended operational horizons, which are acknowledged as limitations of

the current study. The risk of adversarial attacks on DGA sensor streams is discussed further in the deployment section.

Table 8. Robustness analysis under sensor noise injection and missing sensor scenarios (baseline RMSE = 11.42 cycles; Accuracy = 97.34%).

Perturbation Scenario	RMSE (cycles)	Accuracy (%)
No perturbation (baseline)	11.42	97.34
Gaussian noise (5% feature variance)	12.14	96.48
Gaussian noise (10% feature variance)	13.87	94.62
Gaussian noise (20% feature variance)	17.23	91.17
Single gas sensor dropout (zero imputation)	14.81	93.85
Two gas sensor dropout (zero imputation)	18.43	89.24

An analysis of RUL estimation performance across life-stage ranges is presented in Table 9. The model performs best in the near-failure range (RUL < 50 cycles) with RMSE of 7.84 cycles, and performance degrades progressively in the early-life range (RUL > 150 cycles) where gas concentrations carry limited degradation signal. This life-stage dependency is consistent with findings from bearing and battery benchmarks and informs the appropriate deployment context: the model is most reliable and operationally valuable in the mid-to-late life stages when maintenance intervention decisions are most consequential.

Table 9. RUL estimation error by operational life-stage range on the held-out test partition.

Life Stage	RUL Range (cycles)	RMSE (cycles)	MAE (cycles)
Early life	150 – 214	16.37	12.84
Mid life	50 – 150	10.91	7.93
Near failure	0 – 50	7.84	5.62

7. Practical Deployment and Cybersecurity Considerations

7.1 Edge Deployment Feasibility

With approximately 312,000 trainable parameters, the model occupies approximately 4.8 MB of memory in 32-bit floating-point precision, well within the storage capacity of industrial IoT gateway devices. At a mean inference latency of 3.2 ms per 30-step window on CPU and 0.8 ms on GPU, the model suggests feasibility for real-time DGA stream processing at sampling intervals substantially shorter than the hours-to-days typical of online DGA monitoring systems. Post-training quantisation to 8-bit integer precision reduces the model size to approximately 1.2 MB and inference latency to 1.1 ms on ARM Cortex-A72 hardware, with a negligible RMSE increase of 0.34 cycles. These estimates are derived from benchmark measurements on the specified hardware configurations and should be validated through formal field trials on utility-grade edge hardware before operational deployment. The model output can feed a maintenance scheduling module that flags units whose predicted RUL falls below a configurable threshold, with 20 cycles recommended based on the near-failure RMSE analysis in Table 9. Integration with existing SCADA and energy management systems can be accomplished through standard IEC 61850 GOOSE messaging or IEC 60870-5-104 reporting interfaces, though formal integration testing is required before production deployment.

7.2 Cybersecurity and Adversarial Considerations

Deployment in smart grid environments introduces cybersecurity risks that must be explicitly acknowledged. DGA sensor streams transmitted over network interfaces are potential targets for adversarial manipulation, including sensor spoofing, data injection, and replay attacks. An adversary capable of injecting false gas concentration readings could potentially manipulate the model's RUL predictions, either triggering unnecessary maintenance

interventions or masking genuine fault progression. Mitigation strategies include cryptographic authentication of sensor data streams using protocols such as IEC 62351, anomaly detection on the sensor inputs themselves as a pre-processing guard, and cross-validation of DGA readings against physical operating parameters such as load current and winding temperature. The noise robustness analysis in Table 8 provides partial evidence that the model degrades gracefully under random perturbations, but formal adversarial robustness evaluation, including gradient-based adversarial attack testing, is a necessary step before security-sensitive deployment and is identified as a direction for future work. Operational resilience should also account for sensor calibration drift over the deployed lifetime, which differs from random Gaussian noise in that it introduces systematic rather than random bias into the model inputs.

8. Conclusion

This paper presented a hybrid CNN-LSTM deep learning framework for simultaneous remaining useful life estimation and fault classification of power transformers in smart grid networks. Evaluated on a publicly available DGA-based transformer FDD-RUL dataset (Kaggle, 2022), the model couples 1D convolutional layers for local feature extraction with stacked LSTM layers for long-range temporal dependency modelling on 30-step sliding windows of multivariate DGA measurements augmented with IEC 60599-informed gas ratio and trend features. On the held-out test partition, the model achieved RMSE of 11.42 cycles, MAE of 8.31 cycles, R^2 of 0.9621, 97.34% fault classification accuracy, 96.81% macro F1-score, and AUC-ROC values exceeding 0.985 for all four fault classes, reducing RMSE by 21.8% relative to the strongest Bi-LSTM baseline. Ablation studies confirmed the essential contributions of both architectural components and domain-engineered features. Wilcoxon signed-rank tests confirmed that all improvements over recurrent baselines are statistically significant at the 0.05 level. SHAP analysis identified acetylene trend and concentration as the dominant prognostic features, consistent with IEC 60599.

Three practical implications stand out for grid asset managers. First, the multi-task joint modelling approach provides both the categorical fault type and the quantitative remaining life estimate from a single inference pass, reducing the computational overhead of deploying separate diagnostic and prognostic models. Second, post-training quantisation results suggest the model may be deployable on industrial edge hardware with minimal performance penalty, enabling potentially real-time online DGA monitoring without cloud dependency, pending formal field validation. Third, the SHAP feature rankings provide empirical justification for sensor prioritisation in cost-constrained monitoring deployments.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. arXiv:1603.04467. https://arxiv.org/abs/1603.04467

Bustamante, S., Manana, M., Arroyo, A., Castro, P., Laso, A., & Martinez, R. (2021). Dissolved gas analysis equipment for online monitoring of transformer oil: A review. Sensors, 21(12), Article 4058. https://doi.org/10.3390/s21124058

Cao, P., Zhang, S., & Tang, J. (2019). Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning. IEEE Access, 7, 26241–26253. https://doi.org/10.1109/ACCESS.2019.2900295

Cigre. (2003). Guide on transformer intelligent condition monitoring (TICM) systems (Cigre Technical Brochure 227). Cigre.

Duval, M. (2002). A review of faults detectable by gas-in-oil analysis in transformers. IEEE Electrical Insulation Magazine, 18(3), 8–17. https://doi.org/10.1109/MEI.2002.1014963

Guo, P., Fu, J., & Yang, X. (2019). Condition monitoring and fault diagnosis of wind turbines gearbox bearing temperature based on kolmogorov-smirnov test and convolutional neural network model. Energies, 11(9), Article 2355. https://doi.org/10.3390/en11092355

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504–507. https://doi.org/10.1126/science.1127647

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

IEC. (2015). Mineral oil-impregnated electrical equipment in service: Guide to the interpretation of dissolved and free gases analysis (IEC 60599:2015). International Electrotechnical Commission.

Kaggle. (2022). Power transformers FDD and RUL dataset [Data set]. https://www.kaggle.com/datasets/srishtigarg/power-transformers-fdd-and-rul

Kong, Z., Cui, Y., Xia, Z., & Lv, H. (2019). Convolution and long short-term memory hybrid deep neural networks for remaining useful life prognostics. Applied Sciences, 9(19), Article 4156. https://doi.org/10.3390/app9194156

Li, X., Ding, Q., & Sun, J. Q. (2019). Remaining useful life estimation in prognostics using deep convolution neural networks. Reliability Engineering and System Safety, 172, 1–11. https://doi.org/10.1016/j.ress.2017.11.021

Malik, H., Mishra, S., Srivastava, D. C., & Gupta, R. A. (2019). Artificial intelligence-based dissolved gas analysis techniques for fault classification in transformers: A systematic review. IET Generation, Transmission & Distribution, 13(22), 5009–5025. https://doi.org/10.1049/iet-gtd.2019.1094

Mirowski, P., & LeCun, Y. (2009). Dynamic factor graphs for time series modeling. In W. Buntine, M. Grobelnik, D. Mladenic, & J. Shawe-Taylor (Eds.), Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2009), Lecture Notes in Computer Science (Vol. 5782, pp. 128–143). Springer. https://doi.org/10.1007/978-3-642-04174-7_9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Ren, L., Cui, J., Sun, Y., & Cheng, X. (2018). Multi-bearing remaining useful life collaborative prediction: A deep learning approach. Journal of Manufacturing Systems, 43(Part 3), 248–256. https://doi.org/10.1016/j.jmsy.2017.02.013

Shintemirov, A., Tang, W. H., & Wu, Q. H. (2009). Power transformer fault classification based on dissolved gas analysis by implementing bootstrap and genetic programming. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 39(1), 69–79. https://doi.org/10.1109/TSMCC.2008.2007253

Shu, C., Wu, J., Jiang, H., & Zhang, L. (2021). Dissolved gas analysis and deep learning models for transformer fault diagnosis. IET Science, Measurement & Technology, 15(3), 291–303. https://doi.org/10.1049/smt2.12027

Tenbohlen, S., Coenen, S., Djamali, M., Muller, A., Samimi, M. H., & Siegel, M. (2016). Diagnostic measurements for power transformers. Energies, 9(5), Article 347. https://doi.org/10.3390/en9050347

Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN 2017) (pp. 1578–1585). IEEE. https://doi.org/10.1109/IJCNN.2017.7966039

Wu, J., Liao, L., & Li, L. (2020). Remaining useful life prediction with convolutional long short-term memory neural network. In Proceedings of the 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC 2020) (pp. 1–5). IEEE. https://doi.org/10.1109/I2MTC43012.2020.9129483

Zaremba, W., Sutskever, I., & Vinyals, O. (2015). Recurrent neural network regularization. arXiv:1409.2329. https://arxiv.org/abs/1409.2329

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050