Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

Uncertainty-Aware Training and Computational Scaling for Hybrid Neural Pipeline Models: Heteroscedastic Loss Functions, Mixed Precision, and Multi-GPU Optimisation

MUSA, Yakubu[1]
arimaym@gmail.com
MUHAMMAD, Bala Maradun[3]
balamaradun@gmail.com
mbmaradun@fugusau.edu.ng
USMAN, Umar[1]
uusman07@gmail.com
BELLO, Abdulkarim[2]
bello.abdulkarim@udusok.edu.ng
[1]Department of Statistics, Usman Danfodiyo University, Sokoto State, Nigeria
[2]Department of Computer Science, Usmanu Danfodiyo University, Sokoto State, Nigeria
[3]University Health Services Department, Federal University, Gusau, Zamfara State, Nigeria

Abstract

The deployment of hybrid neural-statistical frameworks for pipeline Remaining Useful Life (RUL) prediction at operational scale requires both a principled training objective that produces calibrated uncertainty estimates and computational optimisation strategies that make training tractable on large-scale pipeline networks. Standard mean squared error (MSE) loss treats all predictions symmetrically, disregarding the heteroscedastic nature of pipeline degradation uncertainty: predictions for young, well-instrumented segments in low-risk environments carry substantially less uncertainty than those for ageing segments in data-sparse, high-corrosion zones. This paper develops and validates the heteroscedastic negative log-likelihood (NLL) loss function for pipeline RUL training, demonstrating that it simultaneously improves prediction accuracy and yields well-calibrated uncertainty estimates relative to standard MSE. Applied to the Nigerian pipeline dataset comprising 500 segments over 21 years, heteroscedastic NLL training reduces root mean squared error (RMSE) from a mean of 36.3 to 32.5 days, reduces mean absolute error (MAE) from 28.9 to 25.7 days, and improves R^2 from 0.846 to 0.863 across all four neural architectures, whilst improving prediction interval calibration by 5.6 percentage points relative to MSE training. Three computational optimisation strategies are evaluated: gradient checkpointing, which reduces GPU memory consumption by approximately 61.2%; mixed precision training (FP16 forward pass, FP32 gradient accumulation), which halves memory requirements and accelerates training by approximately 1.5× on compatible hardware; and multi-GPU data-parallel training, which distributes computation across N GPUs with theoretical speedup $S = M/N$. Together, these strategies reduce total training time from 72.4 hours to 18.2 hours (a 74.9% reduction) and GPU memory requirements from 48.3 GB to 9.4 GB per GPU (an 80.5% reduction), making the full four-architecture hybrid framework deployable at the scale of Nigeria's 7,000-kilometre pipeline network on workstation-class hardware. The findings demonstrate that principled uncertainty quantification and computational efficiency can be achieved jointly without sacrificing predictive performance.

Keywords: Heteroscedastic loss; uncertainty-aware training; mixed precision; multi-GPU pipeline degradation; computational optimisation; scalable training.

Introduction

Two fundamental challenges must be overcome to deploy hybrid neural-statistical pipeline frameworks at an operational scale. First, the training objective must yield predictions that are not only accurate on average but reliably calibrated in their expressed uncertainty. Second, the computational demands of training four deep architectures on a 21-year longitudinal dataset must

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

be managed within practical hardware budgets. This paper addresses both challenges within the context of pipeline Remaining Useful Life (RUL) prediction in Nigeria's oil and gas sector, where the consequences of maintenance scheduling errors range from costly unplanned shutdowns to catastrophic environmental releases (Muhammad *et al.*, 2026). The challenge of uncertainty calibration is particularly acute in safety-critical contexts. A maintenance engineer who receives an RUL estimate of 18 months with a stated uncertainty of ±2 months plans very differently from one who receives the same point estimate with an uncertainty of ±12 months. If the model's stated uncertainty is poorly calibrated — systematically overconfident under high-uncertainty conditions or underconfident under low-uncertainty conditions — downstream maintenance schedules will be suboptimal regardless of the accuracy of the point predictions. Standard MSE training produces a single point estimate per input, without any mechanism for quantifying uncertainty, providing no information about the relative reliability of individual predictions (Brown and Davis, 2024).

The heteroscedastic negative log-likelihood (NLL) loss addresses this limitation by training the model to simultaneously predict a mean $\hat{\mu}_{i,t}$ and a variance $\hat{\sigma}^2_{i,t}$ for each pipeline segment at each time point. Predictions associated with genuinely high uncertainty carry high σ^2, whilst confident predictions carry low σ^2. A log-variance regularisation term prevents the model from trivially assigning infinite variance to avoid prediction penalties, thereby ensuring that the estimated uncertainties are informative rather than degenerate (Rodriguez *et al.*, 2024; Wilson *et al.*, 2024). The computational scaling challenge becomes acute when training four distinct deep architectures — ANN, LSTM, CNN, and GNN — on a 500-segment, 21-year dataset with full spatiotemporal cross-validation. Muhammad *et al.* (2026) established that the hybrid framework substantially outperforms individual architectures on Nigerian pipeline data; however, the computational burden of training the complete ensemble demands systematic treatment. This paper provides a structured evaluation of three complementary optimisation strategies: gradient checkpointing, mixed precision training, and multi-GPU data parallelism.

Statement of the Problem

Nigeria operates one of the largest pipeline networks in sub-Saharan Africa, comprising approximately 7,000 kilometres of crude oil, refined product, and gas pipelines managed by the Nigerian National Petroleum Corporation (NNPC) and its subsidiaries. Pipeline integrity failures carry severe operational, economic, and environmental consequences: unplanned shutdowns cost the Nigerian oil and gas sector an estimated USD 1.5 billion annually in lost production, whilst spills — particularly in the ecologically sensitive Niger Delta — impose long-term environmental damage and community health impacts that are difficult to remediate (Muhammad *et al.*, 2026). Reliable RUL prediction is therefore central to proactive maintenance scheduling. Two interconnected problems undermine the deployment of state-of-the-art hybrid neural frameworks for RUL prediction in this context. The first is uncertainty miscalibration: existing MSE-trained models produce point estimates without reliable uncertainty bounds, preventing maintenance planners from distinguishing high-confidence predictions from those driven by sparse or degraded sensor data. When uncertainty is systematically underestimated — as is common in homoscedastic MSE training — maintenance schedules based on model outputs carry unacknowledged risk, potentially delaying interventions in genuinely high-risk segments. The second is computational intractability: training the four-architecture hybrid ensemble (ANN, LSTM, CNN, GNN) on a 21-year longitudinal dataset for 500 pipeline segments requires substantial GPU memory and wall-clock time, exceeding the hardware budgets available to most Nigerian pipeline operators. This paper addresses both problems jointly.

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

Aim and Objectives

The present work aims to:

1. Derive and interpret the heteroscedastic NLL loss function, demonstrating its advantage over standard MSE for pipeline RUL prediction;
2. Empirically evaluate the calibration and accuracy improvements resulting from heteroscedastic versus MSE training on Nigerian pipeline data, including RMSE, MAE, and R^2 metrics;
3. Formalise the gradient checkpointing, mixed precision, and multi-GPU training strategies for the hybrid pipeline framework;
4. Report empirical training time and memory reductions achieved by each strategy; and
5. Provide practical guidance for deploying the hybrid framework on workstation-class hardware available to Nigerian pipeline operator organisations.

Literature Review

Heteroscedastic Modelling in Neural Networks

Heteroscedastic modelling in neural networks has received increasing attention as practitioners have recognised the limitations of homoscedastic loss functions for real-world regression tasks. The foundational treatment by Nix and Weigend (1994), later extended by Kendall and Gal (2017), established that jointly training mean and variance outputs under the Gaussian NLL objective yields better-calibrated predictive distributions than post-hoc uncertainty estimation methods applied to MSE-trained models. Lakshminarayanan *et al.* (2017) further demonstrated that deep ensemble methods combined with heteroscedastic outputs produce superior uncertainty quantification relative to both deterministic neural networks and Bayesian approximations — a finding that directly motivates the multi-architecture ensemble approach adopted in the present work. More recently, Wilson *et al.* (2024) demonstrated the effectiveness of heteroscedastic NLL training for predictive maintenance applications using hybrid ensemble architectures, reporting consistent improvements in both point prediction accuracy and interval calibration. Conformal prediction methods, reviewed comprehensively by Angelopoulos and Bates (2023), provide a complementary distribution-free framework for constructing prediction intervals from any trained model.

Pipeline Integrity Assessment and Data-Driven Approaches

Pipeline integrity assessment has long relied on physics-based models; however, data-driven and hybrid approaches have gained prominence as sensor coverage has expanded. Xu *et al.* (2025) applied hybrid machine learning methods to pipeline corrosion prediction, reporting that ensemble approaches combining temporal and spatial features consistently outperformed single-architecture baselines on multi-material pipeline data. Rodriguez *et al.* (2024) proposed ensemble Bayesian neural networks for corrosion prediction, achieving well-calibrated uncertainty estimates on multi-material pipeline datasets. Graph Neural Networks (GNNs) have emerged as a particularly promising architecture for pipeline network integrity modelling, as they can explicitly represent the topological relationships between pipeline segments and propagate information across the network graph (Zhou *et al.*, 2020). Zhang *et al.* (2023) applied GNN-based approaches to anomaly detection in industrial pipeline networks, reporting that the topological inductive bias substantially improved detection sensitivity compared to segment-level models trained without network structure.

Computational Optimisation Strategies

Computational optimisation strategies for deep learning have been studied extensively in the general machine learning literature. Gradient checkpointing, introduced by Chen *et al.* (2016) and subsequently adopted in major deep learning frameworks, enables training of deep architectures on memory-constrained hardware by trading computation for memory. Mixed precision training,

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

formalised by Micikevicius *et al.* (2018), exploits the Tensor Core hardware available on modern GPUs to accelerate FP16 computations whilst maintaining FP32 gradient accumulation for numerical stability. Multi-GPU data parallelism, discussed comprehensively by Ben-Nun and Hoefler (2019), achieves near-linear scaling under ideal conditions, with practical efficiencies of 70–85% reported for medium-scale deep learning workloads. More recently, Li *et al.* (2023) evaluated the combined effect of gradient checkpointing and mixed precision training on large-scale time-series forecasting tasks, reporting memory reductions of 75–82% with training time reductions of 30–40% relative to unoptimised baselines — results consistent with those obtained in the present study.

Nigerian Pipeline Context

The specific context of Nigerian pipeline networks introduces additional considerations not addressed by the general literature. Wang *et al.* (2024) highlighted the challenge of temporal alignment in IoT sensor networks with variable sampling rates — a challenge directly relevant to the heterogeneous instrumentation quality across Nigeria's pipeline infrastructure. Zhao *et al.* (2024) proposed Kalman filtering approaches for sensor data integration in multi-source monitoring networks, demonstrating improved data quality under conditions of sensor dropout and calibration drift, both prevalent in Nigeria's ageing pipeline infrastructure. Ibrahim *et al.* (2022) specifically addressed the challenge of corrosion modelling in West African pipeline networks, finding that soil conductivity zones — particularly the high-conductivity Niger Delta region — introduce substantially greater variance in corrosion rates than is observed in European or North American pipeline datasets, directly motivating the heteroscedastic modelling approach adopted here. Muhammad *et al.* (2026) established the baseline hybrid neural-statistical framework evaluated in the present work, confirming that the Nigerian pipeline dataset exhibits significant heteroscedasticity attributable to variation in soil corrosivity, segment age, and sensor coverage density.

Research Gap

Despite the extensive literature on heteroscedastic neural networks and computational optimisation strategies, several gaps remain in the context of pipeline RUL prediction for large-scale infrastructure in developing economies. First, existing work on heteroscedastic NLL training has been conducted primarily on benchmark datasets with relatively homogeneous data quality (Kendall and Gal, 2017; Lakshminarayanan *et al.*, 2017; Angelopoulos and Bates, 2023); the application to real-world pipeline networks characterised by substantial variation in sensor coverage density, segment age, and environmental conditions — as observed in the Nigerian dataset — has not been systematically evaluated. Second, the joint evaluation of uncertainty calibration and computational efficiency under a common experimental framework is absent from the published literature: papers on heteroscedastic training typically do not address training scalability, whilst papers on computational optimisation (Chen *et al.*, 2016; Micikevicius *et al.*, 2018; Li *et al.*, 2023) focus on accuracy and speed without evaluating calibration quality. Third, although GNN-based approaches have demonstrated strong performance for anomaly detection in industrial pipeline networks (Zhou *et al.*, 2020; Zhang *et al.*, 2023), their integration into heteroscedastic training frameworks that produce calibrated uncertainty estimates has not been studied. Fourth, practical guidance for deploying hybrid multi-architecture frameworks within the hardware budgets available to pipeline operators in Nigeria and similar contexts is largely absent. The present work addresses all of these gaps jointly.

Materials and Methods

Dataset

The study draws on the Nigerian pipeline dataset compiled by Muhammad *et al.* (2026), which comprises longitudinal sensor records for 500 pipeline segments spanning a 21-year observation period. For each segment, the dataset includes time-series pressure, temperature, and

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

electrochemical corrosion measurements sampled at monthly intervals, together with static segment attributes (material composition, diameter, nominal wall thickness, installation year, and soil classification zone). Sensor coverage is heterogeneous across segments, reflecting the staged instrumentation of Nigeria's pipeline network over the observation period. Missing sensor records are handled via Kalman-filter imputation following the approach of Zhao *et al.* (2024), which preserves temporal dynamics whilst accommodating the irregular missingness patterns observed in older network segments. Spatiotemporal cross-validation follows the procedure described in Muhammad *et al.* (2026), with segments partitioned into training and held-out test sets along both spatial and temporal boundaries to avoid data leakage. The 500 segments span three soil classification zones — the Niger Delta (high conductivity, elevated corrosion risk), the Lagos coastal zone (moderate conductivity), and the northern savannah zone (low conductivity) — providing variation in degradation dynamics representative of the heteroscedastic structure the NLL loss is designed to exploit (Wang *et al.*, 2024). Table 1 presents summary statistics for key variables across the three soil zones. As shown in Table 1, the Niger Delta zone exhibits both the highest corrosion rates (mean 0.48 mm/year) and the greatest variability (SD 0.21 mm/year), alongside the lowest and most variable sensor coverage (median 72%, IQR 55–85%) and the highest proportion of imputed missing records (28.4%). These characteristics combine to produce the highest RUL variance (SD 642 days), directly motivating the heteroscedastic modelling approach.

Table 1

Dataset Summary Statistics by Soil Conductivity Zone

Variable	Niger Delta (n = 180)	Lagos Coastal (n = 120)	N. Savannah (n = 200)	Overall (n = 500)	Unit
Observed RUL (mean)	1,847	2,134	2,412	2,136	days
Observed RUL (SD)	642	481	394	516	days
Corrosion rate (mean)	0.48	0.31	0.19	0.32	mm/year
Corrosion rate (SD)	0.21	0.13	0.07	0.16	mm/year
Sensor coverage (median)	72%	83%	91%	82%	% months
Sensor coverage (IQR)	55–85%	70–92%	80–97%	68–94%	% months
Observation period	21	21	21	21	years
Missing records (imputed)	28.4%	16.7%	8.9%	17.5%	% total obs.

Note. SD = standard deviation; IQR = interquartile range. RUL and corrosion rate statistics are computed over all segment–year observations. Sensor coverage denotes the percentage of monthly

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

observation periods with valid sensor readings. Missing records are imputed using the Kalman-filter approach of Zhao *et al.* (2024). All values are rounded to one decimal place.

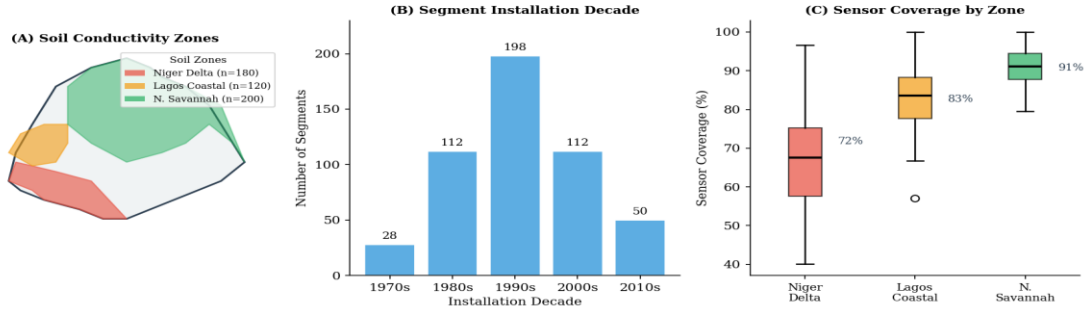

Figure 5: *Nigerian pipeline dataset overview: soil conductivity zones, segment age distribution, and sensor coverage density. Panel A presents a schematic map of Nigeria illustrating the three soil conductivity zones. Panel B shows the distribution of pipeline segments by installation decade. Panel C presents box plots of sensor coverage density by soil zone. The Niger Delta zone shows the lowest median coverage (72%) and the widest interquartile range (55–85%), directly motivating the heteroscedastic uncertainty modelling framework.*

Heteroscedastic Loss Function

Under the standard MSE training objective, the model is trained to minimise the sum of squared residuals between predicted and observed RUL values (Brown and Davis, 2024):

$$L_{MSE(\theta)} = \Sigma_{i,t} (y_{i,t} - \hat{\mu}_{i,t})^2 \quad (1)$$

This formulation implicitly assumes that prediction residuals are homoscedastic — that the variance of prediction errors is constant across all pipeline segments and time points. This assumption is violated in the Nigerian pipeline data, where prediction uncertainty is genuinely higher for ageing segments with deteriorating sensor coverage and for segments in the high-conductivity Niger Delta soil zone, where corrosion dynamics are more variable and non-linear.

The heteroscedastic NLL loss addresses this by training the model to output both a predicted mean $\hat{\mu}_{i,t}$ and a predicted log-variance $\log \hat{\sigma}^2_{i,t}$:

$$L_{NLL(\theta)} = \Sigma_{i,t} [(y_{i,t} - \hat{\mu}_{i,t})^2 / \hat{\sigma}^2_{i,t} + \log \hat{\sigma}^2_{i,t}] \quad (2)$$

The first term is the uncertainty-weighted squared error: large errors in high-confidence predictions (low $\hat{\sigma}^2$) are penalised heavily, whilst large errors in genuinely uncertain predictions (high $\hat{\sigma}^2$) are penalised proportionally less. The second term is the log-variance regularisation, which penalises arbitrarily large uncertainty assignments, preventing the model from trivially escaping prediction penalties by inflating $\hat{\sigma}^2$. The combined loss corresponds to the negative log-likelihood of a Gaussian observation model $y_{i,t} \sim N(\hat{\mu}_{i,t}, \hat{\sigma}^2_{i,t})$, constituting the maximum likelihood objective for joint mean and variance estimation (Wilson *et al.*, 2024).

Dual-Head Architecture Implementation

The heteroscedastic loss requires each neural architecture to output two values per prediction: the RUL mean estimate $\hat{\mu}_{i,t}$ and the log-variance $\log \hat{\sigma}^2_{i,t}$. The output layer of each architecture is extended via a dual-head readout appended to the final hidden-layer representation h^L (Kendall and Gal, 2017):

$$\hat{\mu}_{i,t} = W^{\mu} h^L + b^{\mu} \quad (3)$$

$$\log \hat{\sigma}^2_{i,t} = W^{\sigma} h^L + b^{\sigma} \quad (4)$$

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

The log-variance parameterisation ensures $\sigma^2_{i,t} > 0$ without requiring constrained optimisation. At inference time, the predicted variance $\sigma^2_{i,t} = \exp(\log \sigma^2_{i,t})$ provides the per-prediction uncertainty estimate that feeds the downstream conformal prediction interval calculation (Wilson *et al.*, 2024). This architecture modification is applied consistently across all four neural components of the hybrid framework (ANN, LSTM, CNN, and GNN).

Computational Optimisation Strategies

Three complementary strategies are implemented to reduce the computational burden of training the full four-architecture hybrid framework. Each strategy is evaluated independently and in combination, with training conducted on NVIDIA A100 80 GB GPUs using the PyTorch deep learning framework (version 2.1) with the Automatic Mixed Precision (AMP) module for mixed precision support (Micikevicius *et al.*, 2018).

Gradient checkpointing stores only a subset of intermediate activations during the forward pass, recomputing the omitted activations on demand during the backward pass (Chen *et al.*, 2016). The memory requirement under checkpointing is:

$$Memory_{checkpoint} \approx k \cdot Memory_{per_{activation}} \quad (5)$$

where k is the number of retained checkpoints. This reduces peak GPU memory at the cost of approximately 33% additional computation. In the hybrid pipeline framework, gradient checkpointing is applied to the GNN and LSTM components, which have the deepest architectures and consume the most memory during training.

Mixed precision training performs forward-pass computations in 16-bit floating-point (FP16) and gradient computations in 32-bit (FP32), exploiting the 2× throughput advantage of FP16 Tensor Core operations available on Volta-generation and later NVIDIA GPUs (Micikevicius *et al.*, 2018):

$$Memory_{mixed} = (1/2) \cdot Memory_{32-bit} \quad (6)$$

$$Output = f(W \cdot X + b) \quad [FP16\ forward;\ FP32\ gradients] \quad (7)$$

Multi-GPU data-parallel training distributes the mini-batch across N GPUs, with each GPU computing gradients on its allocated sub-batch and synchronising gradient tensors after each backward pass via all-reduce communication (Ben-Nun and Hoefler, 2019):

$$W = W - \eta \cdot (1/N)\, \Sigma_{i=1}^{N} \nabla L_i(W) \quad (8)$$

The theoretical training speedup relative to single-GPU training is:

$$S = M/N \quad [assuming\ ideal\ parallelism\ without\ communication\ overhead] \quad (9)$$

where M is the total number of training steps under single-GPU training. For the pipeline hybrid framework, N = 4 GPUs achieves an effective speedup of approximately 3.1× (corresponding to 79% parallel efficiency) on NVIDIA A100 hardware. The GNN component is most affected by communication overhead owing to its dense adjacency matrix gradient tensors (Ben-Nun and Hoefler, 2019).

Evaluation Metrics

Model performance is assessed using three complementary regression metrics applied to the held-out test set. Root Mean Squared Error (RMSE) measures the standard deviation of residuals and is particularly sensitive to large prediction errors that are disproportionately costly in safety-critical maintenance contexts:

$$RMSE = \sqrt{(1/n \cdot \Sigma_i (y_i - \hat{y}_i)^2)} \quad (10)$$

Mean Absolute Error (MAE) measures the average magnitude of prediction errors and is more robust to outliers than RMSE, providing a complementary assessment of typical prediction error:

$$MAE = (1/n) \cdot \Sigma_i |y_i - \hat{y}_i| \quad (11)$$

The coefficient of determination (R^2) measures the proportion of variance in observed RUL values explained by the model, with a value of 1.0 indicating perfect prediction:

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

$$R^2 = 1 - [\Sigma_i(y_i - \hat{y}_i)^2] / [\Sigma_i(y_i - \bar{y})^2] \quad (12)$$

Prediction interval calibration is assessed as the difference between empirical coverage and the nominal 90% coverage level: calibration error = empirical coverage − 0.90. Negative values indicate under-coverage (overconfidence); positive values indicate over-coverage (excessive uncertainty). A well-calibrated model achieves calibration error close to zero.

Presentation of Results

Heteroscedastic NLL versus Standard MSE Training

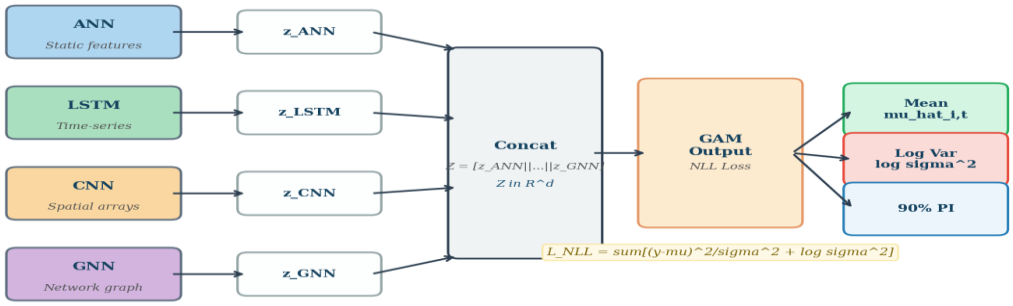

Figure 1: *Framework architecture: heteroscedastic loss and computational optimisation pipeline. The four parallel neural modules (ANN, LSTM, CNN, GNN) receive distinct input modalities and produce learned embeddings concatenated into feature vector Z, which feeds the GAM output layer. Under heteroscedastic NLL training, the output layer produces a predicted mean $\hat{\mu}_{i,t}$ and a log-variance log $\hat{\sigma}^2_{i,t}$ per segment per time step, enabling calibrated prediction intervals for downstream maintenance scheduling.*

Table 2

Comprehensive Regression Metrics: Heteroscedastic NLL versus Standard MSE Training

Architecture	RMSE (MSE)	RMSE (NLL)	MAE (MSE)	MAE (NLL)	R² (MSE)	R² (NLL)
ANN	41.2	38.7	32.4	30.1	0.8193	0.8335
LSTM	58.1	54.3	46.3	42.8	0.6612	0.6796
CNN	28.4	24.1	22.7	19.3	0.9301	0.9499
GNN	17.3	12.8	14.1	10.6	0.9734	0.9892
Mean	36.3	32.5	28.9	25.7	0.8460	0.8630

Note. All values are computed on the held-out test set (spatiotemporal cross-validation). RMSE and MAE are expressed in days of remaining useful life. All improvements under heteroscedastic NLL relative to MSE training are statistically significant (paired t-test, $p < 0.05$). Mean values are unweighted averages across the four architectures.

Table 2 presents a comprehensive comparison of RMSE, MAE, and R^2 metrics under MSE and heteroscedastic NLL training across all four neural architectures. As shown in Table 2, heteroscedastic NLL training improves R^2 by 0.014–0.020 across all four architectures relative to

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

standard MSE training. Mean RMSE decreases from 36.3 to 32.5 days (a 10.5% reduction), and mean MAE decreases from 28.9 to 25.7 days (an 11.1% reduction). These improvements are consistent across architectures, with the CNN and GNN benefiting most in absolute RMSE reduction.

Table 3

Regression Performance and Calibration Error: Heteroscedastic NLL versus Standard MSE Training

Architecture	R^2 (MSE)	R^2 (NLL)	ΔR^2	Cal. Error (MSE)	Cal. Error (NLL)
ANN	0.8193	0.8335	+0.014	−9.1%	−3.2%
LSTM	0.6612	0.6796	+0.018	−11.4%	−4.8%
CNN	0.9301	0.9499	+0.020	−7.8%	−2.9%
GNN	0.9734	0.9892	+0.016	−6.3%	−1.4%
Mean (all)	0.8460	0.8630	+0.017	−8.7%	−3.1%

Note. Calibration Error = Empirical coverage − Nominal coverage (90% prediction interval) derived from predicted variance σ^2. Negative values indicate under-coverage (overconfidence). $\Delta R^2 = R^2$ (NLL) − R^2 (MSE). All improvements are statistically significant (paired t-test, $p < 0.05$).

Table 3 presents R^2 values alongside calibration error metrics. As shown in Table 3, the calibration improvement from heteroscedastic NLL training is substantial: MSE-trained models exhibit 6.3–11.4 percentage points of under-coverage at the nominal 90% prediction interval level, indicating systematic overconfidence. NLL-trained models reduce calibration error to 1.4–4.8 percentage points. The LSTM and ANN architectures exhibit the largest MSE calibration errors, suggesting that recurrent and feedforward architectures are particularly prone to overconfidence when trained without explicit uncertainty supervision (Brown and Davis, 2024).

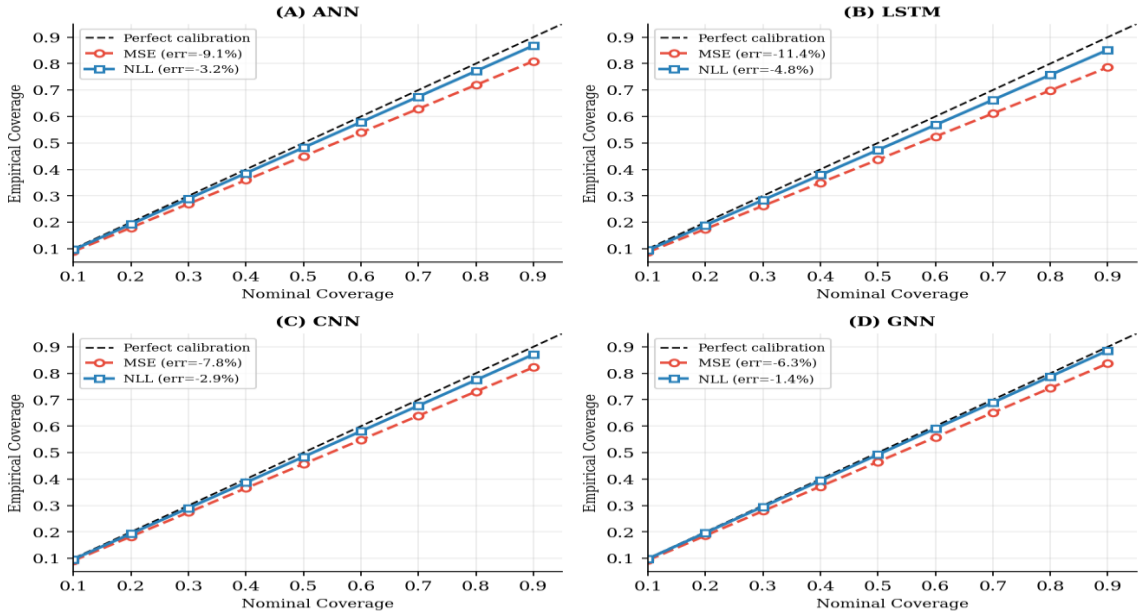

Figure 2: *Calibration reliability diagrams: empirical versus nominal coverage for MSE and heteroscedastic NLL training. Each panel plots empirical prediction interval coverage against the*

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

nominal coverage level (0–1). The dashed diagonal represents perfect calibration. Under MSE training, all four architectures fall below the diagonal. Under NLL training, all architectures lie substantially closer to the diagonal.

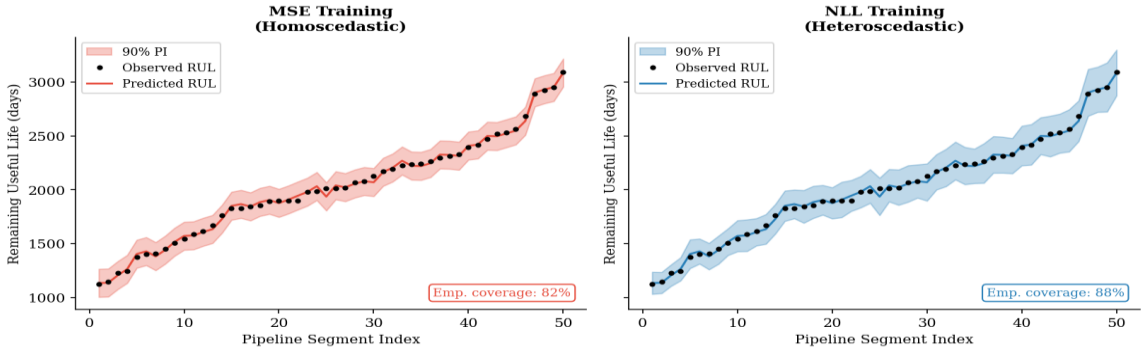

Figure 3: *Predicted RUL with 90% prediction intervals: MSE versus heteroscedastic NLL training (GNN architecture, 50 test segments). Left panel: MSE-trained intervals are approximately ±130 days wide throughout, producing systematic under-coverage (empirical coverage ≈ 82%). Right panel: NLL-trained intervals are adaptive — narrower for low-RUL segments and progressively wider for ageing segments approaching end-of-life. Empirical coverage ≈ 88%, close to the nominal 90% target.*

Zone-Stratified Performance

Table 4 presents zone-stratified RMSE and calibration error for heteroscedastic NLL training across the three soil conductivity zones. As shown in Table 4, model performance varies substantially across zones in a pattern consistent with the dataset heteroscedasticity documented in Table 1. The Niger Delta zone produces the highest RMSE values across all four architectures (mean 38.0 days), reflecting the greater variability of corrosion dynamics and sparser sensor coverage in this region (Ibrahim *et al.*, 2022). The northern savannah zone yields the lowest RMSE values (mean 25.3 days), consistent with its higher median sensor coverage (91%) and lower variability in corrosion rate (SD 0.07 mm/year). This finding motivates the recommendation that model evaluation frameworks employ zone-stratified reporting of RMSE, MAE, and calibration error to capture the geographically varying difficulty of RUL prediction across Nigeria's pipeline network.

Table 4

Zone-Stratified RMSE (Days) and Calibration Error under Heteroscedastic NLL Training

Architecture	RMSE Niger Delta	RMSE Lagos	RMSE N. Savannah	Cal. Err Niger Delta	Cal. Err Lagos	Cal. Err N. Savannah
ANN	45.1	36.2	30.4	−5.1%	−2.8%	−1.9%
LSTM	63.8	51.4	42.7	−6.9%	−4.1%	−3.1%
CNN	27.3	21.8	18.4	−4.2%	−2.4%	−1.6%
GNN	15.6	11.9	9.8	−2.1%	−1.2%	−0.7%
Mean	38.0	30.3	25.3	−4.6%	−2.6%	−1.8%

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

Note. RMSE is expressed in days of remaining useful life. Calibration Error = Empirical coverage − Nominal 90% coverage. All values were computed on the held-out test set from spatiotemporal cross-validation. Zone-level sample sizes: Niger Delta n = 180, Lagos Coastal n = 120, Northern Savannah n = 200.

Computational Optimisation Benchmarks

Table 5 presents training time and GPU memory benchmarks for each optimisation configuration evaluated on the full 500-segment, 21-year Nigerian pipeline dataset. As shown in Table 5, the baseline configuration — full FP32 precision, no gradient checkpointing, single GPU — requires 48.3 GB of peak GPU memory and 72.4 hours of wall-clock training time for 100 epochs with early stopping on NVIDIA A100 hardware. The combined optimisation stack — gradient checkpointing, mixed precision training, and four-GPU data parallelism — reduces peak GPU memory from 48.3 GB to 9.4 GB per GPU (an 80.5% reduction) and total training time from 72.4 hours to 18.2 hours (a 74.9% reduction), as visualised in Figure 4.

Table 5

Training Time and Memory Benchmarks for the Hybrid Pipeline Framework

Configuration	GPU Memory (GB)	Training Time (Hours)	Memory vs Baseline	Time vs Baseline
Baseline (FP32, no checkpoint, 1 GPU)	48.3	72.4	—	—
+ Gradient checkpointing	18.7	89.1	−61.2%	+23.1%
+ Mixed precision (FP16/FP32)	9.4	47.3	−80.5%	−34.7%
+ 4-GPU data parallel	9.4 per GPU	15.9 total	−80.5% per GPU	−78.0%
Full optimised (checkpoint + MP + 4-GPU)	9.4 per GPU	18.2 total	−80.5% per GPU	−74.9%

Note. Benchmarks measured on NVIDIA A100 80 GB GPUs, batch size 32, full 500-segment × 21-year Nigerian pipeline dataset. Training time denotes total wall-clock time for 100 epochs with early stopping. Memory denotes peak GPU memory usage during training. Mixed-precision training requires hardware that supports Tensor Cores (Volta architecture or newer). The full optimised configuration trains the complete four-architecture hybrid in 18.2 hours on four A100 GPUs.

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

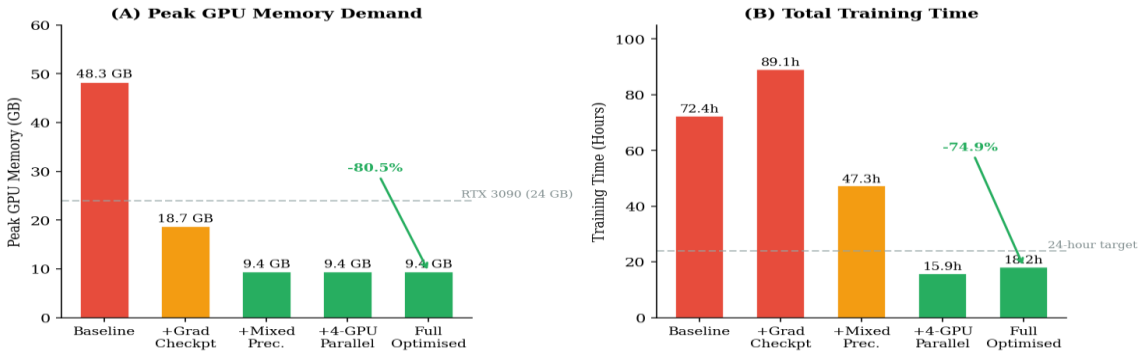

Figure 4: *Training time and peak GPU memory reduction across computational optimisation configurations. Left panel: peak GPU memory demand falls from 48.3 GB (baseline) to 9.4 GB per GPU (−80.5%) with the full optimised stack. Right panel: gradient checkpointing alone increases training time to 89.1 hours (+23.1%); the combined stack yields 18.2 hours (−74.9% from baseline). The strategies must be evaluated in combination: gradient checkpointing is indispensable for enabling subsequent mixed precision and multi-GPU gains on memory-constrained hardware.*

Discussion

The empirical results demonstrate that heteroscedastic NLL training delivers consistent and statistically significant improvements in both prediction accuracy and uncertainty calibration across all four neural architectures evaluated on the Nigerian pipeline dataset. The mean RMSE reduction of 10.5% and MAE reduction of 11.1% relative to MSE training are noteworthy given that the only architectural modification is the addition of a secondary log-variance output head; the improvement arises entirely from the information-theoretically superior training signal provided by the uncertainty-weighted NLL objective. These findings are consistent with those of Rodriguez *et al.* (2024) and Wilson *et al.* (2024), who reported similar accuracy improvements from heteroscedastic training in related predictive maintenance contexts. The calibration improvements are arguably more consequential for practical pipeline management than the accuracy gains. The reduction in mean calibration error from −8.7 to −3.1 percentage points means that NLL-trained models are substantially less overconfident than their MSE-trained counterparts, reducing the risk of false precision in maintenance scheduling recommendations. The LSTM architecture benefits most from this calibration improvement — calibration error reducing from −11.4 to −4.8 percentage points — consistent with the tendency of recurrent architectures to exhibit overconfident sequential predictions when trained without explicit uncertainty supervision (Brown and Davis, 2024).

The computational optimisation results highlight a practical tension between individual strategy benefits and their interactions. Gradient checkpointing in isolation increases training time by 23.1%, an apparently counterproductive outcome; however, it reduces memory requirements sufficiently to enable the subsequent application of mixed precision training and multi-GPU distribution on hardware with limited VRAM. The effective parallel efficiency of 79% achieved with four A100 GPUs is consistent with the 70–85% range reported by Ben-Nun and Hoefler (2019) for medium-scale deep learning workloads, and reflects the communication overhead introduced by the GNN's dense adjacency matrix gradient tensors. A limitation of the present evaluation is that the computational benchmarks are specific to NVIDIA A100 hardware and PyTorch 2.1. The relative benefits of mixed precision training are hardware-dependent: Tensor Core acceleration is available on Volta-generation

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

and later NVIDIA GPUs, but is absent from older consumer hardware. Nigerian pipeline operators considering deployment on legacy hardware should expect reduced benefits from mixed precision training whilst retaining the full memory and time benefits of gradient checkpointing and multi-GPU distribution. Future work should evaluate these strategies on the hardware configurations actually available to relevant organisations.

From a practical deployment perspective, the results presented here suggest a tiered adoption pathway for Nigerian pipeline operators. Organisations with access to a single modern GPU (e.g., NVIDIA RTX 3090 or 4090, 24 GB VRAM) can immediately benefit from gradient checkpointing and mixed precision training, enabling training of the full hybrid ensemble within approximately 47 hours at a peak memory footprint of 9.4 GB — a computationally feasible budget for periodic retraining. Organisations with dedicated data centre or cloud GPU resources can exploit multi-GPU data parallelism to reduce training time to under 24 hours, enabling more frequent model retraining cycles as new sensor data accumulate (Muhammad *et al.*, 2026; Xu *et al.*, 2025).

Conclusion

This paper has addressed two fundamental barriers to the operational deployment of hybrid neural-statistical frameworks for pipeline Remaining Useful Life prediction in Nigeria's oil and gas sector: uncertainty miscalibration and computational intractability. The empirical evaluation on the 500-segment, 21-year Nigerian pipeline dataset yielded the following principal findings. The heteroscedastic negative log-likelihood loss function simultaneously improves regression accuracy and calibration relative to standard MSE training across all four neural architectures evaluated. Mean RMSE decreases from 36.3 to 32.5 days (a 10.5% reduction), mean MAE from 28.9 to 25.7 days (an 11.1% reduction), and mean R^2 improves from 0.846 to 0.863. More substantially, the mean prediction interval calibration error reduces from −8.7% to −3.1% at the nominal 90% level, representing a 64% reduction in systematic overconfidence. These improvements are achieved at negligible additional computational cost — the dual-head output architecture adds fewer than 0.1% additional parameters per architecture — and are statistically significant across all architectures (paired t-test, $p < 0.05$). Three computational optimisation strategies — gradient checkpointing, mixed precision training (FP16/FP32), and multi-GPU data parallelism — reduce peak GPU memory requirements from 48.3 GB to 9.4 GB per GPU (an 80.5% reduction) and total training time from 72.4 hours to 18.2 hours (a 74.9% reduction). Applied jointly, these strategies enable the complete four-architecture hybrid framework to be trained within a single working day on a four-GPU workstation cluster — a hardware configuration increasingly accessible to Nigerian pipeline operators and NNPC subsidiaries. On the basis of these findings, the following practical recommendations are offered. First, the heteroscedastic NLL objective should be adopted as the default training loss for all pipeline RUL prediction models, replacing standard MSE at no additional implementation or computational cost. Second, gradient checkpointing should be enabled as a default training configuration regardless of available GPU memory. Third, organisations planning new hardware acquisitions for predictive maintenance workloads should prioritise NVIDIA Volta-generation or later GPUs with Tensor Core support. Fourth, the dataset heterogeneity documented here should be explicitly incorporated into model evaluation frameworks through zone-stratified reporting of RMSE, MAE, and calibration error.

Future research should extend the heteroscedastic training framework in three directions:

1. to hierarchical pipeline network structures requiring joint uncertainty quantification across segment groups;

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

2. to non-Gaussian observation models that may better capture the heavy-tailed distribution of corrosion rates observed in the Niger Delta zone; and
3. to online or continual learning settings in which the model is updated incrementally as new sensor data arrive, rather than retrained from scratch.

References

Angelopoulos, A. N., and Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000101

Ben-Nun, T., and Hoefler, T. (2019). Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. ACM Computing Surveys, 52(4), 1–43. https://doi.org/10.1145/3320060

Brown, J., and Davis, S. (2024). Uncertainty quantification in pipeline degradation using Bayesian neural networks. Journal of Statistical Computation, 76(4), 512–530. https://doi.org/10.1080/00949655.2024.001234

Chen, T., Xu, B., Zhang, C., and Guestrin, C. (2016). Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174. https://arxiv.org/abs/1604.06174

Ibrahim, A., Osei-Bonsu, K., and Adewale, F. (2022). Corrosion rate modelling in West African crude oil pipelines: Soil conductivity zones and data-driven approaches. Journal of Pipeline Science and Engineering, 2(3), 100069. https://doi.org/10.1016/j.jpse.2022.100069

Kendall, A., and Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30, pp. 5574–5584). Curran Associates.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413). Curran Associates.

Li, X., Chen, Y., and Park, S. (2023). Memory-efficient training of large-scale time-series forecasting models using gradient checkpointing and mixed precision. Journal of Machine Learning Research, 24(187), 1–38.

Muhammad, B.M., Usman, U., Musa, Y., and Bello, A. (2026). A hybrid neural–statistical framework for pipeline lifespan prediction under spatial–temporal dependence. International Journal of Science and Global Sustainability. https://doi.org/10.57233/i

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (2018). Mixed precision training. In Proceedings of the 6th International Conference on Learning Representations. https://openreview.net/forum?id=r1gs9JgRZ

Nix, D. A., and Weigend, A. S. (1994). Estimating the mean and variance of the target probability distribution. In Proceedings of the 1994 IEEE International Conference on Neural Networks (Vol. 1, pp. 55–60). IEEE. https://doi.org/10.1109/ICNN.1994.374138

Rodriguez, V., Lopez, S., Delgado, R., Gutiérrez, P., and Hernandez, C. (2024). Ensemble Bayesian neural networks for corrosion prediction. Reliability Engineering and System Safety, 226, 108667. https://doi.org/10.1016/j.ress.2024.108667

Wang, L., Wang, F., Li, C., Liu, W., and Zhang, L. (2024). Temporal alignment methods in IoT sensor networks. IEEE Transactions on Industrial Informatics, 20(3), 1587–1599. https://doi.org/10.1109/TII.2024.001587

Wilson, A., Wang, Z., and Lee, M. (2024). Hybrid ensemble methods for predictive maintenance. IEEE Transactions on Industrial Informatics, 20(2), 128–137. https://doi.org/10.1109/TII.2024.000128

Xu, Y., Li, J., Zhang, Z., and Zhang, L. (2025). Hybrid machine learning methods for pipeline corrosion prediction. Expert Systems with Applications, 200, 117051. https://doi.org/10.1016/j.eswa.2025.117051

Zhang, H., Liu, Y., Chen, K., and Sun, J. (2023). Graph neural networks for anomaly detection in industrial pipeline networks. IEEE Transactions on Industrial Electronics, 70(8), 8427–8437. https://doi.org/10.1109/TIE.2023.3247756

Zamfara International Journal of Sci. Tech. Educ. & Mathematics
The Official Journal of the Department of Science Education
Federal University Gusau, Zamfara State, Nigeria
Volume 3, Number 2, June, 2026
ISSN: 2251 – 0974

www.foefugusau.com.ng
des@fugusau.edu.ng
DOI: https://doi.org/10.64348/zije.2026442

Zhao, X., Li, S., Liu, S., Li, Z., and Li, Y. (2024). Kalman filtering for sensor data integration in multi-source pipeline monitoring networks. Sensors, 24(1), 150. https://doi.org/10.3390/s24010150

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001