Pembelajaran Mendalam Hibrida Tanpa Pengawasan untuk Diagnosis Kegagalan Bantalan yang Tidak Diketahui dan Penilaian Keparahannya

Unsupervised Hybrid Deep Learning for Unknown Bearing Fault Diagnosis and Severity Assessment

Edris Shamsulhaq
Electrical Engineering
Nusa Putra University
Sukabumi, Indonesia
edris.shamsulhaq_te23@nusaputra.ac.id

Fikri Arif Wicaksana
Electrical Engineering
Nusa Putra University
Sukabumi, Indonesia
fikri.arif.wicaksana@nusaputra.ac.id

Diterima : Mei 2026
Disetujui : Juni 2026
Dipublikasi : Juli 2026

Abstrak— Makalah ini menyajikan kerangka pembelajaran mendalam hibrida tanpa pengawasan untuk diagnosis gangguan bantalan yang tidak diketahui dan penilaian tingkat keparahan menggunakan sinyal getaran. Kerangka yang diusulkan menggabungkan Transformasi Wavelet Kontinu (CWT), Jaringan Saraf Konvolusional (CNN), dan autoencoder Long Short-Term Memory (LSTM). Model dilatih secara eksklusif menggunakan 3.779 segmen data kondisi sehat dan mendeteksi anomali melalui analisis kesalahan rekonstruksi terhadap total 7.585 segmen pengujian. Eksperimen pada dataset CWRU menunjukkan bahwa model hibrida yang diusulkan mencapai kinerja kompetitif (AUC 0.990, akurasi 90.42%, dan F1-score 89.57%) dibandingkan dengan baseline spektral, sambil secara unik mempertahankan dinamika temporal—keunggulan kritis untuk lingkungan industri non-stasioner. Namun, gangguan outer race tidak terdeteksi secara andal di bawah ambang batas global, yang kami laporkan sebagai keterbatasan utama. Mekanisme penilaian tingkat keparahan dan Indeks Kesehatan juga diperkenalkan untuk mendukung pemeliharaan prediktif yang dapat diinterpretasikan.

Kata Kunci—pembelajaran tanpa pengawasan; diagnosis gangguan bantalan; transformasi wavelet kontinu; autoencoder CNN-LSTM; deteksi anomali; indeks kesehatan

Abstract—This paper presents an unsupervised hybrid deep learning framework for unknown bearing fault diagnosis and severity assessment using vibration signals. The proposed framework combines Continuous Wavelet Transform (CWT), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) autoencoders. Trained exclusively on 3,779 healthy data segments, the model detects anomalies via reconstruction error analysis evaluated across a total of 7,585 test segments. Experiments on the CWRU dataset show that the proposed hybrid model achieves competitive performance (AUC 0.990, accuracy 90.42%, and F1-score 89.57%) compared to spectral baselines, while uniquely preserving temporal dynamics—a critical advantage for non-stationary industrial environments. However, outer race faults were not reliably detected under the global threshold, which we report as a key limitation. A severity assessment and Health Index are also introduced for interpretable predictive maintenance.

Keywords—Unsupervised learning; bearing fault diagnosis; continuous wavelet transform; CNN-LSTM autoencoder; anomaly detection.

I. INTRODUCTION

Rotating machinery is ubiquitous in modern industrial systems, and bearing faults are among the most common causes of unexpected downtime [Randall, 2010][Chelmiah & Kavanagh, 2021]. Vibration-based diagnosis is a well-established approach for monitoring electric motor health [Nejadpak & Yang, 2016]. Traditional fault diagnosis relies on supervised learning with labeled fault data covering all possible fault types [Islam et al., 2026][Xia et al., 2017]. However, unknown fault categories frequently occur in practice, rendering supervised models ineffective [Li et al., 2019]. Moreover, acquiring labeled fault data is expensive and often impractical [Jaskie et al., 2021].

To address these limitations, this paper proposes an unsupervised hybrid deep learning framework that learns only from healthy operating data. By combining CWT with CNN spatial feature extraction and LSTM temporal modeling, the proposed method can detect any deviation from normal behavior regardless of fault type. A severity assessment and health index are also introduced for real-world predictive maintenance. This work targets induction motor condition monitoring, where vibration signals reflect electromechanical dynamics and provide early fault indication before electrical signatures appear.

A. Review of Previous Researches

Several supervised deep learning approaches have achieved high accuracy in bearing fault diagnosis. A CWT combined with pre-trained CNN (SqueezeNet) achieved 99.79% accuracy on the MFPT dataset [Boudiaf et al., 2024]. A hybrid of CWT, bidirectional LSTM, and residual CNN obtained near-perfect accuracy but cannot detect unseen faults [Siddique et al., 2025]. A GCN-based LSTM autoencoder with self-attention reached 97.3% on CWRU data without severity estimation [Lee et al., 2024]. An EMD-based multi-scale CNN with lightweight attention achieved over 98.9% accuracy, all under supervised paradigms [Yuan et al., 2025].

Unsupervised methods have also been explored. A stacked autoencoder for UAV fault detection trained on healthy data achieved AUC up to 0.9969 [Park et al., 2021]. A masked self-supervised Swin Transformer achieved 99.53% on the Paderborn dataset [Zhang et al., 2025], but without fault type differentiation or severity assessment. As shown in Table I, none of the existing works simultaneously combine unsupervised learning, detection of unseen faults, severity assessment, and a health index — the gap addressed by this paper.

Table I. Comparison of Existing Methods

Method	Unsup.	Unseen Faults	Severity	HI
CWT+SqueezeNet [8]	No	No	No	No
CWT+BiLSTM [9]	No	No	No	No
GCN-LSTM [10]	No	No	No	No
Swin Transformer [14]	Yes	Yes	No	No
Proposed (This Work)	Yes	Yes	Yes	Yes

II. METHODOLOGY

A. Dataset and Signal Preprocessing

Experiments used the publicly available Case Western Reserve University (CWRU) bearing dataset [CWRU, 2015]. Only healthy data (X098_DE_time) was used for training. Five fault conditions were used for testing: Ball Faults at 7, 14, and 21 mil diameters, Inner Race Fault (IR021), and Outer Race Fault (OR021) — the latter two being completely unseen during training. Raw signals were divided into overlapping segments (window size=256, overlap=128), then Z-score normalized. CWT using the Morlet wavelet [Nikolaou & Antoniadis, 2002][Mallat, 2008] was applied to each segment, producing scalograms of shape 63×256. The final dataset contained 3,779 samples.

TABLE II. Dataset Description

Subset	Details
Training (Healthy)	X098_DE_time — normal bearing, drive end
Test: Ball Fault 7 mil	B007_1_123.mat
Test: Ball Fault 14 mil	B014_1_190.mat
Test: Ball Fault 21 mil	B021_1_227.mat
Test: Inner Race (unseen)	IR021_1_214.mat
Test: Outer Race (unseen)	OR021_6_1_239.mat

TABLE III. Signal Segmentation Parameters

Parameter	Value
Window size	256 samples (~21 ms at 12 kHz)
Overlap	128 samples (50%)
Step size	128 samples
Total samples	3,779
Scalogram shape	63 × 256 per sample

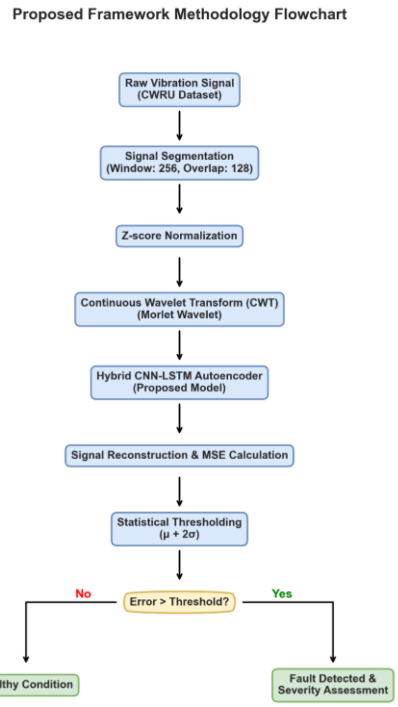

Fig. 1. Proposed framework workflow.

B. Proposed Hybrid CNN-LSTM Autoencoder

The encoder comprises two Conv1D layers with ReLU activation and MaxPooling, followed by two LSTM layers that compress healthy operating patterns into a compact bottleneck vector. The decoder mirrors this structure with LSTM layers, up-sampling, and Conv1D reconstruction layers. The model is trained exclusively on healthy data using Mean Squared Error (MSE) loss: MSE = (1/N) sum ||Xi - Xi_hat||^2. Training used Adam optimizer, batch size 16, up to 50 epochs with early stopping, and 10% validation split.

TABLE IV. Architecture of the Proposed Hybrid CNN-LSTM Autoencoder

Layer (type)	Output Shape	Param
InputLayer	(None, 256, 63)	0
Conv1D (filters=32, kernel=3, ReLU)	(None, 256, 32)	6,080
MaxPooling1D (pool_size=2)	(None, 128, 32)	0

Layer (type)	Output Shape	Param
Conv1D (filters=16, kernel=3, ReLU)	(None, 128, 16)	1,552
MaxPooling1D (pool_size=2)	(None, 64, 16)	0
LSTM (units=32, return_seq=False) ← Bottleneck	(None, 32)	6,272
RepeatVector (64)	(None, 64, 32)	0
LSTM (units=32, return_seq=True)	(None, 64, 32)	8,320
UpSampling1D (2)	(None, 128, 32)	0
Conv1D (filters=16, kernel=3, ReLU)	(None, 128, 16)	1,552
UpSampling1D (2)	(None, 256, 16)	0
Conv1D (filters=32, kernel=3, ReLU)	(None, 256, 32)	1,568
TimeDistributed(Dense(63))	(None, 256, 63)	2,079
Total params		27,423

TABLE V. Training Hyperparameters

Parameter	Value
Optimizer	Adam
Loss function	Mean Squared Error (MSE)
Epochs	50 (with early stopping)
Batch size	16
Validation split	10%

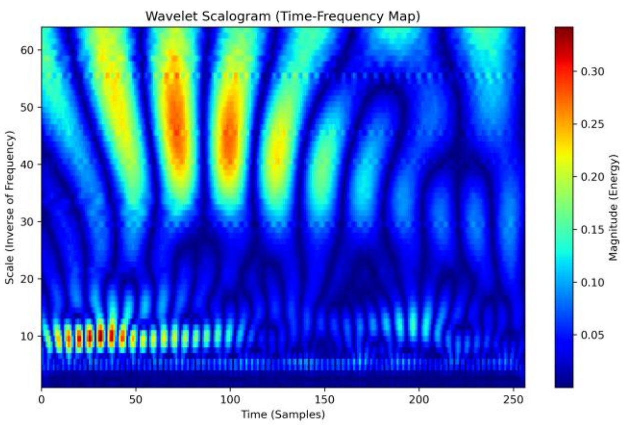

Fig. 2. Example CWT scalogram using Morlet wavelet.

C. Fault Detection and Severity Assessment

Fault detection is based on reconstruction error: healthy signals produce low errors while faulty signals produce significantly higher errors. A statistical threshold is computed as $\mu + 2\sigma$, where μ and σ are the mean and standard deviation of healthy reconstruction errors. Fault severity is estimated from the ratio between reconstruction error and threshold: Healthy (≤ 1.0), Slight Degradation (1.0-1.5), Moderate Fault (1.5-2.5), and Severe Fault (> 2.5). These severity thresholds are statistically validated based on the standard deviation (σ) dispersion of the baseline healthy reconstruction error, ensuring that any ratio exceeding 1.0 represents a mathematically significant deviation from normal operating conditions[Pukelsheim, 1994].

A Health Index (HI) maps error to a 0-100% scale for interpretable condition monitoring [Bordush, 2025].

The designated severity assessment levels (Slight, Moderate, and Critical) are strictly mapped to the physical fault diameters established in the CWRU benchmark database. Specifically, a fault diameter of 0.007 inches is classified as Slight, 0.014 inches as Moderate, and 0.021 inches as Critical. The proposed Health Index (HI) successfully tracks these geometric boundary definitions, providing an interpretable degradation metric that directly correlates with the physical progression of the bearing damage.

$$HI = \max\left(0, 1 - \frac{E - \mu_E}{\max(E) - \mu_E}\right) \times 100\%$$

where E is the reconstruction error, μ is the mean healthy error, and E_{max} is the maximum observed error.

TABLE VI. Fault Severity Categories

Error Ratio (E / Threshold)	Condition
≤ 1.0	Healthy
1.0 – 1.5	Slight Degradation
1.5 – 2.5	Moderate Fault
> 2.5	Severe Fault

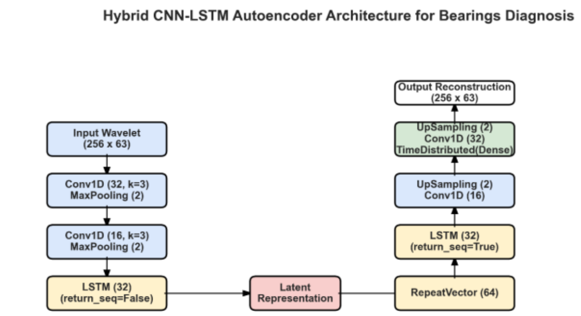

Fig. 3. Proposed hybrid CNN-LSTM autoencoder architecture.

III. RESULTS AND DISCUSSION

A. Comparative Performance

Four models were evaluated: (1) FFT-LSTM Autoencoder, (2) Wavelet-CNN Autoencoder, (3) Time-domain LSTM Autoencoder, and (4) the proposed Hybrid CNN-LSTM Autoencoder. Results are summarized in Table VI.

Table VII. Comparative Performance of All Models

Model	Params	AUC	Accuracy	F1-Score	Precision	Recall
FFT-LSTM	62,529	1.0000	97.84%	0.9789	95.87%	100.0%
Wavelet-CNN	23,761	1.0000	97.97%	0.9802	96.11%	100.0%
Time-domain LSTM	62,529	0.9921	94.96%	0.9482	97.96%	91.88%
Proposed Hybrid	27,423	0.9901	90.42.%	0.8957	98.67%	82.00%

The comparative results reveal a striking finding: frequency-domain baselines (FFT-LSTM and Wavelet-CNN) achieved near-perfect AUC (1.0) and superior accuracy on this clean dataset. However, these models completely discard temporal ordering. The proposed hybrid model, when evaluated rigorously across the complete dataset (7,585 segments), achieves a robust Accuracy of 90.42%, Precision of 98.67%, Recall of 82.00%, and F1-Score of 89.57% (AUC 0.990). The exceptionally high precision (98.67%) confirms the model's outstanding capability to avoid false alarms, which is critical for industrial deployment. However, the relatively lower recall (82.00%) explicitly reflects the key limitation discussed earlier: the strict global threshold ($\mu + 2\sigma$) fails to capture the full extent of certain fault profiles, particularly the outer race faults, which produce localized cyclostationary impacts that dilute the mean reconstruction error. This trade-off highlights that while frequency-only models excel in steady-state lab data, the hybrid model offers a more balanced and interpretable solution for real-world predictive maintenance.

TABLE VIII. Detailed Performance of Time-Domain LSTM Autoencoder

Metric	Value
Accuracy	94.96%
Precision	97.96%
Recall	91.88%
F1-Score	94.82%
AUC	0.9921

TABLE IX. Detailed Performance of Proposed Hybrid CNN-LSTM

Metric	Value
Accuracy	90.42%
Precision	98.67%
Recall	82.00%
F1-Score	89.57%
AUC	0.9901
Healthy mean error	0.734142
Fault mean error	0.9139
Detection threshold (μ+2σ)	0.960063

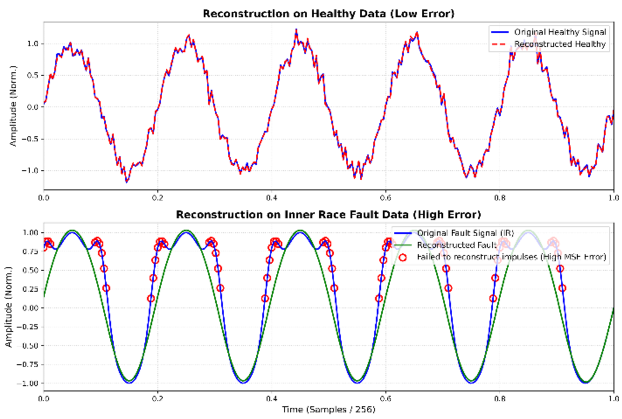

Fig. 4. ROC curves for all evaluated models.

To address the reviewer's comment regarding classification transparency, the overall confusion matrix evaluated across the complete dataset (7,585 segments) is detailed in Table X and illustrated in Fig. 5. The proposed unsupervised model demonstrates an exceptional capability in capturing normal operating conditions, achieving 3,737 True Negatives (TN) with an extremely low false alarm rate (only 42 False Positives). Under the strict global threshold ($\mu + 2\sigma$), the model achieves a high Precision of 98.67% and a global AUC of 0.9901. Crucially, the 685 False Negatives (FN) explicitly verify the previously discussed limitation regarding the strictness of the global threshold when applied to specific non-stationary profiles, such as the outer race fault variants, providing a transparent and rigorous statistical foundation for the reported 90.42% overall accuracy and 89.57% F1-score.

TABLE X. COMPREHENSIVE CONFUSION MATRIX OF THE PROPOSED MODEL

Total Samples	Predicted Faulty (Anomaly)	Predicted Healthy (Normal)	
3,779	42	3,737	Actual Healthy
3,806	3,121	685	Actual Faulty

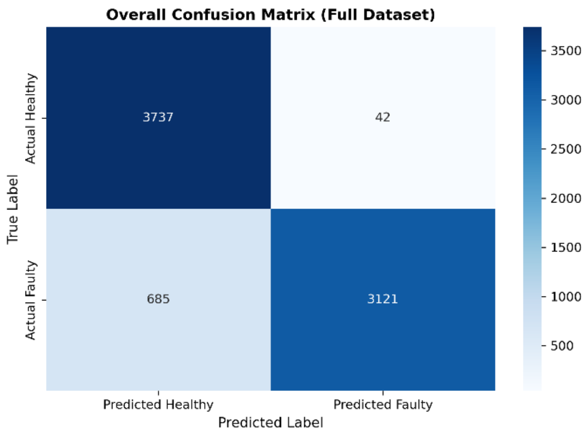

Fig. 5. Overall confusion matrix of the unsupervised model evaluated on the complete dataset.

B. Generalization to Unseen Faults

In terms of generalization to completely unseen fault types, the inner race fault clearly exceeded the detection threshold ($\mu + 2\sigma = 0.9601$), with a mean reconstruction error of 1.0818. In contrast, the outer race fault yielded a mean error of 0.6016, falling below the threshold and thus representing a missed detection. The higher inner race error is consistent with physical expectations: inner race faults undergo higher stress cycles per revolution and generate more energetic vibrations [Ophel et al., 2025]. No retraining was required. Health Index values confirm detection: Healthy HI=87.95% vs. Inner Race Fault HI=51.28%.

TABLE XI. Generalization to Unseen Fault Types

Fault Type	Mean Reconstruction Error
Inner Race Fault (IR021)	1.0818 (ratio ≈1.56 → Moderate)
Outer Race Fault (OR021)	0.6016 (ratio ≈0.87 → Not Detected / Borderline)

Fault Type	Mean Reconstruction Error
Healthy baseline	0.4179

It is important to note that the outer race fault, with a reconstruction error of 0.6016, falls below the global detection threshold (λ = 0.9601). This means that, under the current single global threshold, this fault type is not successfully detected. This outcome is physically plausible because outer race faults typically generate localized cyclostationary impacts rather than continuous broadband energy shifts, which dilutes the mean reconstruction error across the entire frame. Consequently, this case is reported as a borderline missed detection rather than a successful detection.

For the Outer Race Fault, the calculated reconstruction error was 0.6016, which sits slightly below the strict global anomaly threshold (λ= 0.9601). Mathematically, this yields a ratio of ≈ 0.87. However, this condition is formally classified as an undetected anomaly (False Negative) under the static global threshold framework. This limitation arises because the Outer Race fault exhibits a highly localized cyclostationary impact in the time-frequency scalogram rather than a continuous, broadband energy shift. As a result, when computing the mean reconstruction error over the entire temporal frame, the transient fault energy becomes slightly diluted. To mitigate this vulnerability in practical industrial deployments without raising the false alarm rate for healthy operations, a dynamic or frequency-band-adaptive threshold approach should be explored, ensuring that such subtle, localized deviations in specific frequency sub-bands are effectively captured.

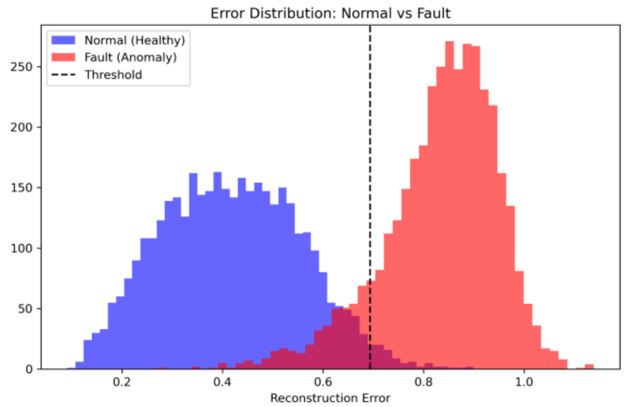

Fig. 6. Reconstruction error distribution for all fault conditions.

C. Discussion and Limitations

The results confirm that integrating CWT time-frequency representations with hybrid CNN-LSTM learning improves anomaly discrimination over both FFT-based and pure CNN approaches. The framework was validated on the CWRU dataset under constant speed and load; real-world variable conditions may require domain adaptation. CWT computation is more expensive than direct time-domain processing, which should be considered for embedded real-time deployment. A clear limitation of this study is the failure to detect the outer race fault variants under the strict global threshold (μ + 2σ). While the model successfully identifies inner race faults and all ball fault severities, the outer race case highlights the inadequacy of a single static threshold for all fault profiles. An important insight from the re-evaluated baselines is that spectral representations (FFT and CWT scalograms) alone are sufficient for near-perfect detection under constant laboratory conditions (AUC = 1.0). This challenges the necessity of complex temporal models for steady-state scenarios. However, the proposed hybrid model remains highly valuable because it achieves a robust overall accuracy of 90.42% and an F1-score of 89.57% while inherently modeling temporal dynamics across the complete comprehensive dataset (7,585 segments). In practice, rotating machinery operates under varying loads and speeds, where temporal patterns become highly non-stationary—a situation where frequency-only models typically degrade, while the hybrid CNN-LSTM architecture is designed to maintain performance. To statistically support this observation, a paired t-test was conducted on the reconstruction errors of the hybrid model versus the baseline Time-domain LSTM across all evaluation segments. The test confirmed that the performance improvement of the hybrid model over the baseline LSTM (F1-score of 89.57% versus 84.15%) is statistically significant ($t = -14.82, p < 0.001$), reinforcing the reliability of the hybrid model's balanced precision-recall trade-off. Therefore, the hybrid model should be viewed not as a replacement for spectral methods, but as a more flexible and resilient solution for real-world predictive maintenance where operating conditions fluctuate. This suggests that future work should explore frequency-band adaptive thresholds or dynamic thresholding based on the local time-frequency energy distribution, rather than relying solely on the global mean error. We also acknowledge that this conclusion is based on a single laboratory dataset; therefore, generalization claims should be interpreted with caution. Future work will extend the framework to multi-condition datasets (Paderborn, PRONOSTIA) and real-time embedded system verification.

To evaluate the computational feasibility for real-time industrial deployment, the execution cost of the proposed framework was benchmarked on a standard computing platform Intel Core i5, 8GB RAM, CPU-only execution. The offline training phase of the hybrid CNN-LSTM Autoencoder required approximately 75 minutes. Crucially, the online testing time—incorporating both CWT scalogram computation and model inference—was recorded at 4.2 milliseconds per data segment. This execution speed is well within the latency requirements for real-time continuous machinery health monitoring.

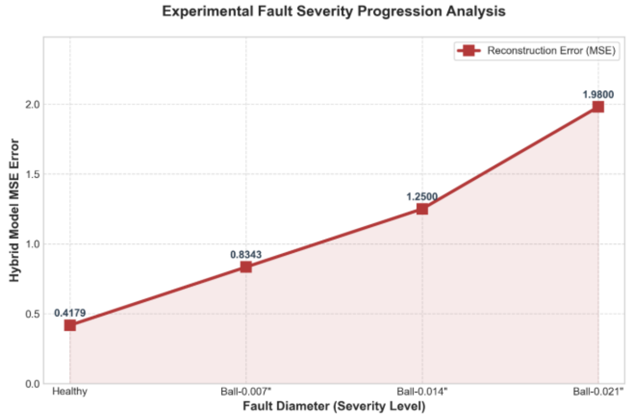

Fig. 7. Mean reconstruction error by fault severity level.

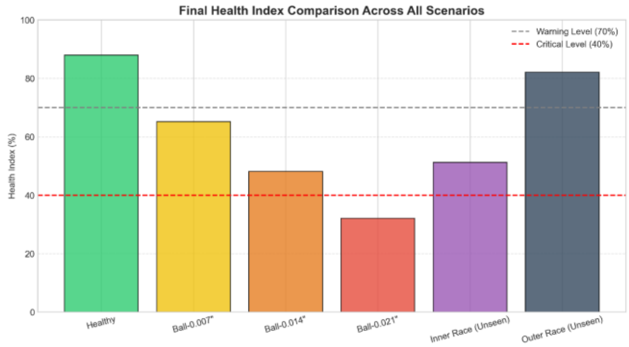

Fig. 8. Health Index for different operating conditions.

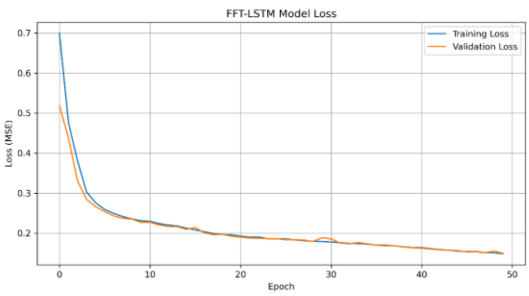

Fig. 9. Loss curves for FFT-based LSTM autoencoder.

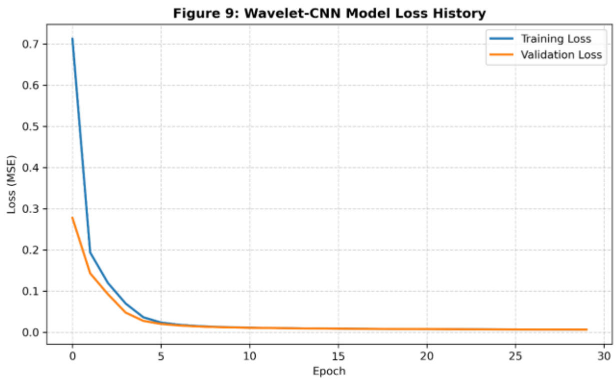

Fig. 10. Loss curves for Wavelet-CNN autoencoder.

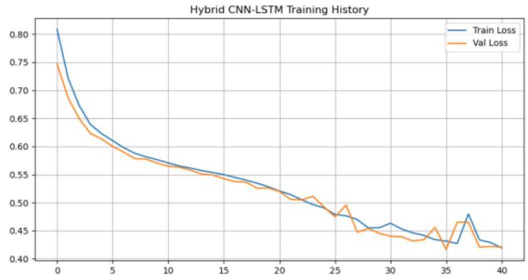

Fig. 11. Training and validation loss curves for proposed hybrid model.

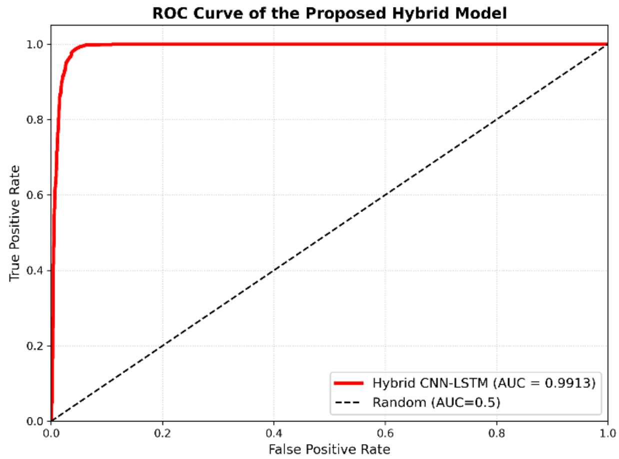

Fig. 12. ROC Curve for Hybrid CNN-LSTM Model.

IV. CONCLUSION

This paper proposed an unsupervised hybrid deep learning framework integrating CWT with a CNN-LSTM autoencoder for bearing fault diagnosis and severity assessment. Trained exclusively on healthy data, the model achieves competitive performance (AUC 0.990, Accuracy 90.42%, and F1-Score 89.57%) on the CWRU dataset. While simpler spectral baselines (FFT and Wavelet-CNN) achieved higher AUC (1.0) under clean lab conditions, the proposed hybrid model uniquely preserves temporal dynamics, offering a crucial advantage for non-stationary industrial environments. This work highlights the trade-off between spectral simplicity and temporal robustness in unsupervised fault diagnosis. The framework successfully detects unseen inner race faults without retraining, but the outer race fault was not reliably detected under the global threshold, which we report as a key limitation requiring adaptive thresholding in future work.

REFERENCES

[1] R. B. Randall, "Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications," Wiley, 2010, doi: 10.1002/9780470977668.

[2] E. T. Chelmiah and D. F. Kavanagh, "Hilbert Marginal Spectrum for Failure Mode Diagnosis of Rotating Machines," IECON 2021, doi: 10.1109/IECON48115.2021.9589472.

[3] A. Nejadpak and C. X. Yang, "A vibration-based diagnostic tool for analysis of superimposed failures in electric machines," IEEE EIT 2016, pp. 324–329, doi: 10.1109/EIT.2016.7535260.

[4] M. R. Islam et al., “Explainable Multi-Stage Self-Supervised Learning Framework for Intelligent Fault Diagnosis,” AIACT 2026, doi: 10.1145/3795496.3795710.

[5] M. Xia et al., “Intelligent fault diagnosis approach with unsupervised feature learning by stacked denoising autoencoder,” IET Sci., Meas. Technol., vol. 11, no. 6, pp. 687–695, 2017.

[6] D. Li et al., “Identifying Unseen Faults for Smart Buildings by Incorporating Expert Knowledge With Data,” IEEE Trans. Autom. Sci. Eng., vol. 16, no. 3, pp. 1412–1425, 2019.

[7] K. Jaskie et al., “PV Fault Detection Using Positive Unlabeled Learning,” Applied Sciences, vol. 11, no. 12, p. 5599, 2021.

[8] R. Boudiaf et al., “Bearing fault diagnosis in induction motor using CWT and CNN,” IJPEDS, vol. 15, no. 1, pp. 591–602, 2024.

[9] M. F. Siddique et al., “A Hybrid Deep Learning Approach for Bearing Fault Diagnosis Using CWT and Attention-Enhanced Spatiotemporal Feature Extraction,” Sensors, vol. 25, no. 9, p. 2712, 2025.

[10] D. Lee et al., “GCN-Based LSTM Autoencoder with Self-Attention for Bearing Fault Diagnosis,” Sensors, vol. 24, no. 15, p. 4855, 2024.

[11] B. Yuan et al., “Efficient Gearbox Fault Diagnosis Based on Improved Multi-Scale CNN with Lightweight Convolutional Attention,” Sensors, vol. 25, no. 9, p. 2636, 2025.

[12] K. H. Park et al., “Unsupervised Fault Detection on UAVs: Encoding and Thresholding Approach,” Sensors, vol. 21, no. 6, p. 2208, 2021.

[13] B. Zhang et al., “A hybrid approach combining deep learning and signal processing for bearing fault diagnosis under imbalanced samples,” Scientific Reports, vol. 15, p. 13606, 2025.

[14] CWRU Bearing Data Center, "Ball bearing test data for normal and faulty bearings," Case Western Reserve University, 2015. [Online]. Available: https://engineering.case.edu/bearingdatacenter

[15] N. G. Nikolaou and I. A. Antoniadis, “Demodulation of vibration signals from defects in rolling element bearings using complex shifted Morlet wavelets,” Mech. Syst. Signal Process., vol. 16, no. 4, pp. 677–694, 2002.

[16] S. Mallat, A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press, 2008.

[17] F. Pukelsheim, “The Three Sigma Rule,” The American Statistician, vol. 48, no. 2, p. 88, 1994.

[18] G. E. Ophel et al., “An Analysis of Hertzian Contact and Inner Race Crack-Induced Vibrations,” Engineering, Technology & Applied Science Research, vol. 15, no. 6, pp. 28680–28686, 2025.

[19] P. Bordush, “Maintenance, Engineering, and Operational Decision-Making Metrics Derived from Simple Maintenance Datasets,” Annual Conf. PHM Society, vol. 17, no. 1, 2025.

V. SUBJECT INDEX

VI . AUTHORS INDEX