Advances in Engineering Innovation Vol.17 Issue 9 EWA Publishing
Available Online: 19 August 2026 DOI: 10.54254/2977-3903/2026.36264

A reliable rolling bearing fault diagnosis method based on Titan

Bingcong Li

Dipont KCS Chengdu, Chengdu, China

2069648300@qq.com

Abstract. The accuracy of fault diagnosis for rolling bearings degrades sharply when operating conditions shift. Existing high-precision classifiers often experience a drop of over 50% in predictive accuracy when speed or load fluctuates, which seriously jeopardizes the reliability of industrial equipment health monitoring. Titan, a recently proposed architecture for long-context language modelling, tackles a similar challenge of maintaining performance across varying contexts. TitanDiag adapts this mechanism to fault diagnosis. The underlying rationale is that a persistent memory accumulates evidence across operating conditions and stabilizes predictions when the current segment alone is ambiguous. The architecture places Titan's dual-path memory (a long-term store gated by surprise plus a short-term First-In-First-Out (FIFO) buffer) inside a Transformer encoder. The multi-view front-end provides three complementary representations for every vibration segment, namely the raw waveform, the Fourier magnitude spectrum, and the continuous wavelet transform scalogram. At inference, Monte Carlo dropout produces per-prediction uncertainty scores that align naturally with Titan's surprise metric. On the CWRU and PU bearing benchmarks, TitanDiag attains 99.25% accuracy on the challenging PU-C2 low-speed condition, where TimeMachine and TSCMamba drop to 41.68% and 60.20%, respectively. The mean error-detection Area under the Receiver Operating Characteristic Curve (AUROC) reaches 0.9668, well above the best baseline of 0.9391, demonstrating that the memory-driven variance inflation produces uncertainty estimates that are closely aligned with actual misclassification patterns.

Keywords: rolling bearing fault diagnosis, Titan memory architecture, cross-condition generalization / domain shift multi-view signal encoding, uncertainty estimation / Monte Carlo dropout

1. Introduction

Rolling element bearings make up about 40–50% of failures in rotating machinery—more than any other drivetrain component [1, 2]. In most cases, unexpected bearing failure bears economic cost and safety risks. These stakes have made the automated fault diagnosis of bearings one of the core concerns of prognostics and health management, garnering effort for years. In particular, deep learning has progressively reformed bearing diagnosis [3].

Convolutional networks got there first. Hierarchical time-frequency features can indeed separate different fault types apart [4, 5]. A recurrent architecture then learns temporal dependencies across successive shaft revolutions [3]. Transformers took a different road, breaking the bottleneck of the temporal stream with pairwise self-attention that captures everything at once [6]. Attention was later hooked onto convolutional backbones in hybrid models like MB-ViT and Swin-T, capturing global dependency across segments. The most recent turn, though, is pointing towards structured state-space models [7, 8]. Mamba is the champion of this turn, swapping quadratic self-attention for linear-time state-space dynamics [9]. There are three instantiations now charting the frontier of sequential diagnosis: TSCMamba performs feature scanning along both temporal and channel axes [10]. TimeMachine stacks four consecutive Mamba blocks to achieve multi-scale temporal modeling [11]. xLSTM replaces the scalar memory cell of vanilla LSTMs with matrix-valued states and exponential gating mechanisms [12].

None of the existing models in this paradigm preserve historical diagnostic features beyond the current vibration segment. Every single prediction starts from scratch. This assumption holds fine while training and test conditions match. When operating conditions vary, such as varying shaft speeds and reduced loads, feature segments well generalized under the training distribution become ambiguous, leaving the model unable to leverage cross-condition diagnostic evidence. The PU dataset clearly validates this limitation [13]. Under the low-speed C2 working condition, TSCMamba's accuracy drops sharply from approximately 99% to 59.73%, while TimeMachine performs even worse at only 41.68%. Both models have training data available. Their performance degradation stems from the inability to transfer diagnostic evidence across different operating condition domains.

We address this gap by adapting Titan's surprise-gated long-term memory to bearing diagnosis [14]. Titan's neural memory stores key-value pairs beyond the attention window, writing selectively based on surprise scores. We propose TitanDiag, which embeds Titan's dual-path memory (long-term store plus short-term FIFO buffer) within a Transformer encoder. A multi-view frontend provides three complementary feature representations, including time-domain waveforms, Fourier magnitude spectra, and continuous wavelet transform (CWT) scalograms, while the memory module aggregates cross-condition diagnostic evidence. This design allows the model to retrieve and condition on past informative patterns when faced with novel operating conditions, outperforming conventional Monte Carlo dropout in uncertainty quantification performance.

2. Method

2.1. Problem formulation

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be a dataset of vibration segments, where $x_i \in R^L (L = 2048)$ and $y_i \in \{1, \ldots, C\}$ is the fault class (C = 10 for CWRU, C = 8 for PU). Segments span K operating conditions with distinct speed–load parameters. A standard classifier $f_\theta : R^L \rightarrow \Delta^{C-1}$ treats each x_i independently—no information flows between segments. This paper hypothesizes that maintaining a persistent memory M that accumulates diagnostic patterns across conditions and is queried at inference can stabilise predictions under condition shift. The Titan architecture provides a viable framework: selective storage gated by surprise metric, dual-path retention (long-term and short-term), and attention-based retrieval [14]. The challenge is adapting these token-level mechanisms to fixed-dimensional vibration representations. For this task, the structure of the proposed method is shown in Figure 1.

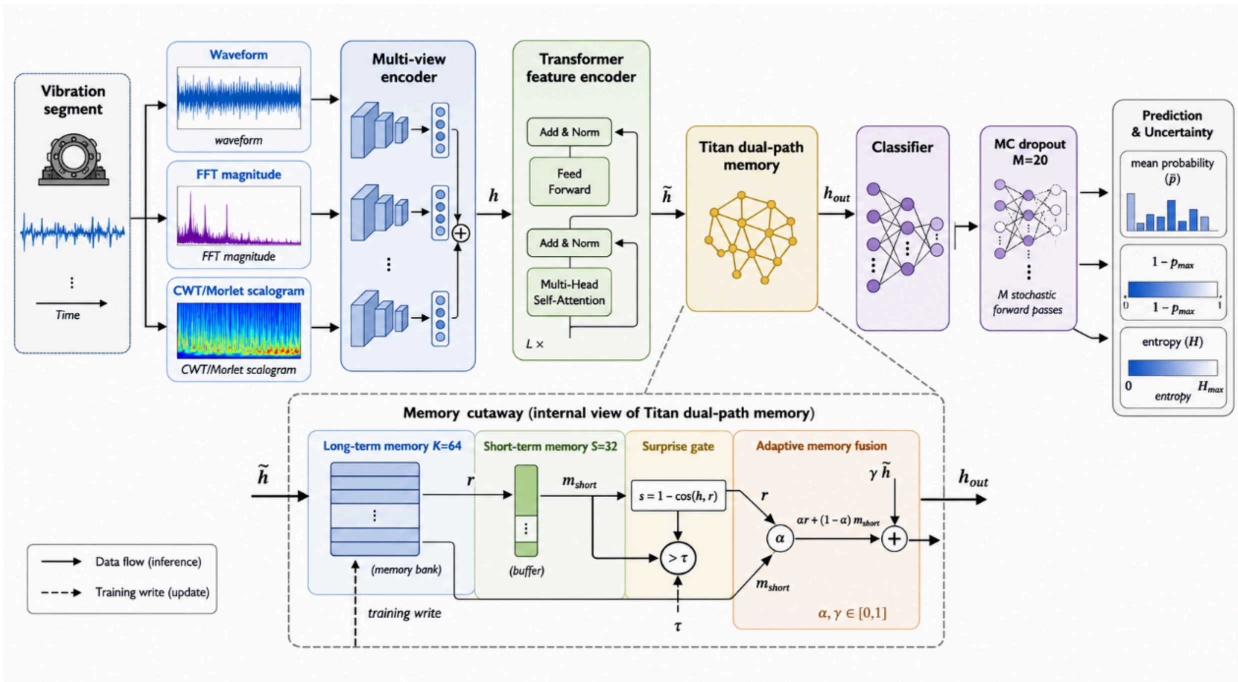

Figure 1. The structure of the proposed method

2.2. Multi-view signal encoding

A single representation cannot capture all diagnostic information. The raw waveform preserves impulsive fault morphology but obscures characteristic frequencies (BPFO, BPFI, BSF). The Fourier magnitude spectrum reveals periodic components yet discards temporal localization—transient impulses and sustained tones at the same frequency become indistinguishable. Time-frequency analysis resolves this ambiguity. We extract three complementary views from each x :

$$v_t = x \in \mathbb{R}^L \tag{1}$$

$$v_f = \left| \mathscr{F}(x) \right|_{1:\lfloor L/2 \rfloor} \in R^{\lfloor L/2 \rfloor} \tag{2}$$

$$v_{tf} = \left| W_\psi(x) \right| \in R^{S \times L} \tag{3}$$

where F is the DFT, W_ψ the CWT with mother wavelet ψ , and $S = 128$.

We utilize the complex Morlet wavelet $\psi(t) = \pi^{-1/4} e^{i\omega_0 t} e^{-t^2/2}$ ($\omega_0 = 6$). Its Gaussian envelope achieves optimal time-frequency localisation, while the complex output retains phase information useful for distinguishing faults with overlapping spectra. Scales are logarithmically spaced from 1 to 128, covering structural resonances at the high end down to shaft-rate modulations; linear spacing would crowd samples at high frequencies and starve the mid-range where BPFO, BPFI, and BSF reside.

Each view is processed by a dedicated 1D convolutional encoder. The time-domain and frequency-domain encoders share an identical architecture. Layer k computes:

$$h^{(k)} = GELU(BN(Conv1D(h^{(k-1)}; W^{(k)}, k^{(k)}))) + Proj(h^{(k-1)}) \tag{4}$$

where $Proj$ is a 1×1 convolution aligning dimensions when the residual path changes channel count or temporal resolution. The time-frequency encoder follows the same structure but with 2D convolutions to exploit the joint time-scale plane. View-specific features are concatenated and projected: $h = W_{fuse}[h_t; h_f; h_{tf}] + b_{fuse} \in \mathrm{R}^D$, where D = 256.

2.3. Transformer context encoder

Bearing fault signatures are temporally structured: the spacing between consecutive impulses encodes the fault frequency. To capture this within-segment context, we pass h through a single Transformer encoder layer. Sinusoidal positional encoding is injected before self-attention. The resulting Transformer output $\widetilde{h}$ serves as the input of the subsequent neural memory module.

2.4. Titan neural memory module

The module receives $\widetilde{h} \in R^D$ and returns $h_{out} \in \mathrm{R}^D$, maintaining a long-term memory $M_{long} \in \mathrm{R}^{K\times D}$ (K = 64) and a short-term FIFO buffer $M_{short} \in \mathrm{R}^{S\times D}$ (S = 32).

Long-term retrieval. Given query $\widetilde{h}$, a soft combination of stored entries is computed via scaled dot-product attention:

$$r = \sum_{j=1}^{K} \alpha_j M_{long}[j], \alpha_j = \frac{exp((\widetilde{h}W_q)\cdot(M_{long}[j]W_k)/\sqrt{D})}{\sum_{l=1}^{K} exp((\widetilde{h}W_q)\cdot(M_{long}[l]W_k)/\sqrt{D})} \tag{5}$$

with $W_q, W_k \in \mathrm{R}^{D\times D}$.

Surprise and selective write. The cosine distance between query and retrieval defines the surprise score:

$$s = 1 - \frac{\widetilde{h}\cdot r}{\|\widetilde{h}\|\|r\|} \in [0, 2] \tag{6}$$

When $s > \tau$ (learnable threshold, initialised at 0.5), the least recently used slot $j^* = argmin_j t_j$ is overwritten: $M_{long}[j^*] \leftarrow \beta \cdot M_{long}[j^*] + (1-\beta)\cdot \widetilde{h}$, with $\beta = 0.95$. We replace the original gradient-norm-based surprise metric in vanilla Titan with cosine distance calculation. This design eliminates per-sample backpropagation overhead during inference while retaining the core screening logic: highly similar historical patterns produce $s \approx 0$, while unseen novel feature inputs drive s close to the upper bound 2.

We set K = 64. Ten fault classes across four conditions could generate roughly 40 distinct patterns; 64 slots accommodate intra-class variation. Reducing to K = 32 cost 0.4% accuracy on PU-C2; K = 128 added nothing.

Short-term buffer. Not every diagnostic signal clears the surprise threshold. An incipient inner-race defect may produce a single impulse per revolution—detectable but not different enough from background to trigger a write. A FIFO buffer of S = 32 recent Transformer outputs catches these transient events regardless. The readout applies two-head attention:

$$m_{short} = MHA_{short}(query = \widetilde{h}, key = B, value = B) \tag{7}$$

where $B \in \mathrm{R}^{S\times D}$ holds the buffer contents and $d_k = D/2 = 128$. At 12 kHz with 2048-sample windows, 32 segments span roughly eight shaft revolutions at 1,500 rpm—long enough to catch multiple fault impulses without retaining stale data after a condition change.

Retrieved memories are fused by a learned gate. A two-layer MLP outputs the mixing weight.

3. Results

3.1. Datasets and preprocessing

Two publicly available bearing datasets were selected. The CWRU dataset acts as a widely adopted benchmark with mild operating condition variations [15]; PU provides deliberate, large condition shifts [13].

CWRU contains 12kHz drive-end vibration from an SKF6205-2RS deep-groove ball bearing [15]. Single-point faults were introduced via electro-discharge machining at the inner race, the outer race (6 o'clock position), and the ball. Three fault diameters were used (0.18, 0.36, and 0.54mm), giving 9 fault classes plus a healthy baseline, and data were collected under four motor loads (0, 1, 2, and 3hp) corresponding to shaft speeds of 1797--1730~rpm. We split signals from every condition-class pair into 2048-point sliding windows with 50% overlap, generating roughly 1000 signal samples for each fault class under each operating condition. The data were partitioned into training, validation, and test sets (60/10/30), stratified by both class and operating condition.

PU was designed for more demanding condition-shift evaluation [13]. The Paderborn University test rig uses a 425W permanent-magnet synchronous motor driving a shaft on 6203 rolling bearings. Eight classes are defined: one healthy state and seven damaged states—outer-ring fatigue pitting (KA05, KA07), outer-ring EDM notches (KA05, KA07), inner-ring EDM notches (KI01, KI03), and an inner-ring drilled hole (KI07). Vibration signals were collected under three separate operating profiles. C1 (1500 rpm, 1 kN radial load, 0.7 Nm torque) acts as the baseline reference case; C2 reduces the rotational speed to 900 rpm while keeping load and torque constant; C3 reverts to 1500 rpm with torque decreased to 0.1 Nm. The 40% speed gap between C1 and C2, with all other parameters fixed, serves as the primary cross-condition robustness test. Signals were down-sampled to 4kHz, divided into non-overlapping 2048-sample windows, and split 60/10/30 with the same stratified scheme used for CWRU.

Every window was normalised to zero mean and unit variance before feature extraction. FFT magnitude spectra retained the first 1024 positive-frequency bins. The continuous wavelet transform employed a complex Morlet wavelet at 128 logarithmically spaced scales. All reported results are means and standard deviations over five independent training runs, each initialised with a different random seed.

3.2. Baseline methods

Nine baselines cover the methodological spectrum from classical machine learning to state-space architectures. XGBoost extracts handcrafted DIT-FFT frequency features and conducts classification via gradient-boosted decision trees, acting as the non-deep-learning baseline lower bound [16]. ResNet-18 (1D) adapts residual learning to one-dimensional wavelet-transformed vibration signals [17]. QCNN augments convolutional layers with quantum-inspired feature maps that expand representational capacity beyond classical convolutions [18]. Swin-T and MB-ViT represent the Transformer-based frontier: Swin-T applies shifted-window self-attention to time-frequency representations, while MB-ViT fuses MobileNet-style inverted bottleneck convolutions with a Vision Transformer backbone [7, 8]. Transform1D provides a pure multi-head self-attention baseline without memory, convolution, or recurrence [19]. xLSTM extends the classical LSTM with matrix-valued memory cells and exponential gating, addressing vanishing-gradient limitations while retaining sequential processing [12]. TimeMachine chains four Mamba structured state-space blocks for linear-complexity temporal modelling [11]. TSCMamba extends Mamba with dual-path temporal and channel-wise scanning and represents the strongest state-space diagnostic competitor to TitanDiag [10]. All aforementioned baseline models are re-implemented under a unified experimental evaluation framework based on PyTorch 2.0. Hyperparameters were tuned on the CWRU validation split and transferred unchanged to PU, replicating the condition-shift scenario where TitanDiag cross-condition memory should matter most.

3.3. Classification accuracy

Table 1 reports per-condition accuracy across the seven test configurations. On the four CWRU conditions (0–3 hp), nearly every deep model exceeds 98%, with TitanDiag at 99.95–100.00%. These conditions differ only in motor load, shifting shaft speed by roughly 70 rpm—a modest domain gap that modern architectures handle without difficulty. The CWRU results confirm that TitanDiag does not sacrifice within-distribution accuracy for its memory capabilities.

Table 1. Comparison of the proposed method with baselines

Model	CWRU-0	CWRU-1	CWRU-2	CWRU-3	PU-C1	PU-C2	PU-C3
XGBoost	91.18 ± 3.27	89.98 ± 0.54	86.66 ± 1.17	91.20 ± 0.66	77.68 ± 1.72	66.00 ± 2.91	67.23 ± 2.61
ResNet	99.88 ± 0.15	99.84 ± 0.12	99.66 ± 0.68	99.90 ± 0.20	99.95 ± 0.11	96.02 ± 0.69	99.33 ± 0.78
QCNN	94.56 ± 2.38	97.32 ± 1.71	98.02 ± 2.48	96.80 ± 5.13	94.33 ± 3.77	95.75 ± 1.85	96.20 ± 2.84
xLSTM	97.10 ± 2.33	99.46 ± 0.39	99.96 ± 0.05	99.50 ± 0.00	98.50 ± 0.00	92.75 ± 3.10	98.08 ± 0.47
Transform1D	95.66 ± 4.28	93.86 ± 0.38	92.20 ± 1.51	93.64 ± 1.16	95.85 ± 0.32	87.50 ± 1.86	93.40 ± 1.28
Swin-T	99.06 ± 0.46	99.48 ± 0.27	99.94 ± 0.09	99.98 ± 0.04	93.15 ± 3.27	94.20 ± 0.65	99.69 ± 0.07
MB-ViT	97.00 ± 0.28	98.06 ± 1.74	99.44 ± 0.43	98.10 ± 2.49	96.98 ± 1.21	88.17 ± 1.87	96.75 ± 1.90
TimeMachine	84.40 ± 2.88	89.18 ± 1.70	91.86 ± 0.87	94.56 ± 1.09	68.12 ± 4.13	42.42 ± 2.01	70.50 ± 0.88
TSCMamba	98.46 ± 2.38	99.36 ± 0.23	99.20 ± 0.28	99.06 ± 0.52	87.13 ± 0.38	60.20 ± 1.16	86.18 ± 1.22
TitanDiag	99.95 ± 0.04	99.98 ± 0.03	100.00 ± 0.00	100.00 ± 0.00	99.98 ± 0.03	99.25 ± 0.18	99.93 ± 0.05

The PU dataset provides a sharper stress test. PU-C1 (1500 rpm, 1 kN) and PU-C3 (1500 rpm, 0.1 Nm) both yield high baselines; the torque reduction in C3 mainly changes signal amplitude scaling, an effect that normalization layers can mostly compensate for. PU-C2 (900 rpm) cuts shaft speed by 40% relative to C1. This speed reduction shifts fault-characteristic frequencies, alters impulse spacing within each 2048-sample window, and changes the modulation patterns that CNNs and attention mechanisms have learned to recognise at the training speed.

Models lacking cross-segment historical memory suffer drastic accuracy degradation under the PU-C2 low-speed working condition. TimeMachine falls to 42.42% (a drop exceeding 50 pp from C1). TSCMamba drops to 60.20%. xLSTM retains 92.75%, its gated recurrent structure implicitly smoothing sequential representations. Among CNN-based models, ResNet reaches 96.02%—the best baseline—because 1D convolutions provide partial robustness to frequency shifts through translation invariance. QCNN and Swin-T hold at 95.75% and 94.20%, respectively. TitanDiag reaches 99.25% on PU-C2, a margin of 3.23 pp over ResNet and 39–57 pp over the state-space models. The memory module lifts accuracy from strong-but-brittle to robust: retrieved long-term patterns provide a conditioning signal that stabilises the classifier when the current window deviates from the training distribution.

4. Discussion

The PU-C2 results isolate the mechanism behind TitanDiag's robustness. The 39–57 percentage point margin over TimeMachine and TSCMamba cannot be attributed to encoder capacity or attention depth—ablation shows that replacing the long-term memory module with a plain state-space block erases nearly the entire gain, while removing the short-term buffer only destabilises transient fault signatures. This ablation outcome verifies that persistent long-term memory serves as a necessary enabler for stable cross-condition generalization.

Multi-view encoding is necessary but not sufficient. Its removal causes the largest single ablation drop, confirming that memory requires representational richness to operate. Yet richness alone is fragile: without persistent storage, the same multi-view features degrade sharply when input statistics shift. The asymmetry is clear—memory converts representational richness into robustness, but richness without memory does not endure distribution change.

5. Conclusion

TitanDiag closes the cross-condition accuracy gap in bearing fault diagnosis through a surprise-driven long-term memory mechanism, rather than relying on deeper encoders or richer features alone. On the demanding PU benchmark, it outperforms the strongest state-space competitor by 39–57 percentage points under condition shift, with ablation confirming memory as the dominant driver of both accuracy and AUROC.

The practical value lies in calibrated uncertainty: an error-detection AUROC of 0.9668 enables reliable human-in-the-loop delegation, minimising silent misclassification risk. Notably, the memory–surprise coupling contributes disproportionately to uncertainty quality—its removal degrades AUROC by 0.0423, which represents a higher relative loss than the corresponding 2.15-point accuracy drop—suggesting that knowing when the model is uncertain is more dependent on memory than knowing what the fault is.

Limitations remain. Ablation studies validated Transformer depth (N = 4), but window length (2048 samples) and wavelet scale range (1-128) were fixed by preliminary search and interact non-trivially with memory dynamics: longer windows reduce the write-to-memory frequency but may miss rapid transients; shorter windows improve temporal resolution at the cost of a lower per-segment signal-to-noise ratio. Systematic co-optimization of these sensing parameters with the memory architecture is left for future work.

References

[1] Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing, 25*(2), 485–520.

[2] Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical Systems and Signal Processing, 42*(1–2), 314–334.

[3] Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing, 138*, 106587.

[4] Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2017). A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors, 17*(2), 425.

[5] Li, X., Zhang, W., & Ding, Q. (2019). Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal Processing, 161*, 136–154.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30*.

[7] Xiao, G., Yao, J., Zhong, L., Xiao, Z., & Lu, X. (2025). MB-ViT: MBConv vision transformer with time-frequency feature fusion for bearing fault diagnosis. *Neural Computing and Applications, 37*, 22801–22825.

[8] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on*

Computer Vision (pp. 10012–10022).

[9] Ahamed, M. A., & Cheng, Q. (2025). Mamba-based architectures for time series analysis: A survey. *Information Fusion*.

[10] Ahamed, M. A., & Cheng, Q. (2025). TSCMamba: Mamba meets multi-view learning for time series classification. *Information Fusion, 120*, 103079.

[11] Ahamed, M. A., & Cheng, Q. (2024). TimeMachine: A time series is worth 4 Mambas for long-term forecasting. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)* (pp. 1688–1695).

[12] Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., et al. (2024). xLSTM: Extended Long Short-Term Memory. In *Advances in Neural Information Processing Systems, 37*.

[13] Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *Proceedings of the European Conference of the Prognostics and Health Management Society*.

[14] Behrouz, A., Zhong, P., & Mirrokni, V. (2024). Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*.

[15] Case Western Reserve University Bearing Data Center. (n.d.). 12k drive end bearing fault data. *Case Western Reserve University*.

[16] Xiang, C., Ren, Z., Shi, P., & Zhao, H. (2021). Data-driven fault diagnosis for rolling bearing based on DIT-FFT and XGBoost. *Complexity, 2021*, 4941966.

[17] Liang, P., Wang, W., Yuan, X., Liu, S., Zhang, L., & Cheng, Y. (2022). Intelligent fault diagnosis of rolling bearing based on wavelet transform and improved ResNet under noisy labels and environment. *Engineering Applications of Artificial Intelligence, 115*, 105269.

[18] Li, Y., Wang, X., Liu, Z., & Zhang, H. (2022). Quantum-inspired convolutional neural networks for bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement, 71*, 1–12.

[19] Ding, X., & He, Q. (2017). Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement, 66*(8), 1926–1935.