Physics-Guided Uncertainty Quantification for Trustworthy Bearing Fault Diagnosis

Mahmudul Hasan Rohan, MD. Ashikur Rahman Shah, and MD. Waliuzzaman Talha
Department of Industrial and Production Engineering
Jashore University of Science and Technology
Jashore, Bangladesh

***Abstract*—Bearing faults in induction motors cause unplanned downtime in continuous manufacturing sectors, where unpredicted motor failure propagates directly into production delays. Existing deep learning diagnostic models optimize for accuracy but provide no uncertainty estimate, making it impossible to distinguish a confident correct prediction from a confident wrong one. We present PhysicsAE, combining three components: (1) a physics-guided adaptive frequency mask derived from SKF 6205-2RS bearing geometry that reduces autoencoder validation loss by 55.9% over a standard MSE baseline at identical model capacity; (2) MC Dropout uncertainty quantification producing per-prediction confidence scores without additional parameters; and (3) the Physics-Uncertainty Triage (PUT) Framework, a three-zone deployment model. Evaluated on the CWRU benchmark, the PUT Framework achieves Zone 3 (confident-wrong) $= 0\%$ across all threshold settings in cross-load evaluation and reduces Zone 3 from 41.3% to 39.6% in cross-speed evaluation. Five-seed validation confirms the fault-severity uncertainty ordering holds in 5/5 seeds with standard deviation <0.0001. An ablation confirms that physics-guided training and encoder-embedded Dropout are jointly necessary: removing either collapses uncertainty to zero.**

***Index Terms*—Bearing fault diagnosis, physics-informed learning, uncertainty quantification, MC Dropout, trustworthy AI, autoencoder.**

I. Introduction

Predictive maintenance for rotating machinery relies on models that classify bearing fault conditions from vibration signals. In high-volume manufacturing sectors, such as the textile and garment industries, continuous motor operation makes early fault detection economically critical [1], [2]. However, classification accuracy alone is insufficient for safe factory deployment. A model with no uncertainty mechanism treats every prediction identically—a wrong prediction carries the same apparent authority as a correct one. In shift-based production environments, operators act on diagnostic alerts without independent verification, making confident-wrong predictions the highest-risk failure mode.

Current autoencoder-based methods use reconstruction error as an anomaly score [3] but are largely deterministic: a single forward pass produces one output with no confidence estimate. Physics-informed approaches encode domain knowledge into training objectives [6] but similarly provide no uncertainty bounds. Neither line of work adequately addresses the deployment question: *when should an operator trust a diagnosis?*

This paper addresses that gap with three contributions:

- A physics-guided adaptive frequency mask that encodes bearing fault frequencies into the reconstruction loss, reducing validation loss by 55.9% at identical model capacity.
- MC Dropout uncertainty quantification embedded in the encoder, producing fault-severity-ordered confidence scores without extra parameters, validated across five independent random seeds.
- The PUT Framework, a three-zone deployment triage that achieves Zone $3 = 0\%$ in cross-load evaluation and reduces dangerous misdiagnoses in cross-speed evaluation.

II. Related Work

Physics-informed fault detection. Zhang et al. [5] designed WDCNN with a wide first kernel to capture fault harmonics. Alam et al. [6] encode BPFO/BPFI priors into CNN loss functions. Both improve accuracy but provide no uncertainty estimates and no deployment triage mechanism.

Uncertainty-aware diagnosis. Gal and Ghahramani [4] established MC Dropout as a practical Bayesian approximation. Zhou et al. [11] apply Bayesian deep learning to fault diagnosis under noise but do not connect uncertainty to operational deployment zones. Kabir et al. [12] survey neural network-based uncertainty quantification in engineering but focus broadly on methodologies rather than specific operational deployment zones.

Trustworthy deployment frameworks. Zhou et al. [7] propose a probabilistic framework for trustworthy bearing diagnosis and argue that reliability quantification is essential for safety-critical deployment. Our PUT Framework operationalises this argument through a concrete three-zone triage validated under two independent protocols.

Gap. No prior work simultaneously encodes bearing physics into the reconstruction loss *and* produces deployment-oriented uncertainty estimates validated across multiple generalisation protocols and random seeds.

III. Methodology

A. Dataset and Preprocessing

The CWRU dataset [8] uses a SKF 6205-2RS Drive-End bearing at 12 kHz. Four conditions are studied: Normal, InnerRace (IR), OuterRace (OR), BallFault (BF), each at four loads (0–3 HP; 1730–1797 RPM). Signals are segmented using a sliding window of 1,024 samples (85.3 ms) with 75% overlap (stride 256) using a Hann window prior to FFT. Each window is z-score normalised, then clipped to the training-set 1st/99th percentile and rescaled to $[-1, 1]$; clipping prevents impulse-dominated windows from distorting the min-max range. The balanced dataset contains 1,888 windows per class. Splits are *file-level*: all windows from a given recording file go entirely to one partition, eliminating overlap leakage between train and test. Final split: 70% train / 15% validation / 15% test, stratified by class, seed 42.

B. Model Architecture

PhysicsAE uses a symmetric 1D convolutional autoencoder. The encoder comprises three `Conv1D--BatchNorm--GELU` blocks (channels 1→32→64→128, stride 2), each followed by `Dropout`(p=0.15), and a fully-connected 64-dimensional bottleneck. The decoder mirrors this with `ConvTranspose1D` blocks and `tanh` output activation. **Both PhysicsAE and BaselineAE share 2,258,049 parameters.** Dropout layers add zero parameters. Any performance difference is therefore attributable solely to the physics constraint and stochastic inference. Mean inference latency for T=30 MC passes is 18 ms per sample on a Tesla T4 GPU, suitable for online monitoring at sensor polling rates of $\leq$50 Hz.

C. Physics-Guided Adaptive Frequency Mask

For the CWRU SKF 6205-2RS bearing (N=9, d=7.940 mm, D=39.040 mm, α=0°):

$$\text{BPFO} = 3.5848\, f_s, \quad \text{BPFI} = 5.4152\, f_s, \quad \text{BSF} = 2.3567\, f_s \tag{1}$$

where f_s=$n/60$ Hz. For each $f_k \in$ {BPFO, BPFI, BSF, FTF, f_s}, the adaptive mask $M(f, n) \in [b, 1]$ over the one-sided rFFT axis (513 bins, 0–6 kHz) is:

$$M(f, n) = b + (1{-}b) \sum_{k=1}^{K} \exp\!\left(-\frac{(f - k \cdot f_k)^2}{2\sigma^2}\right) \tag{2}$$

with K=5 harmonics per fault frequency, baseline weight b=0.1, and σ=15 Hz. The value σ=15 Hz exceeds the frequency resolution Δf=12000/1024=11.7 Hz, ensuring each harmonic bump spans at least one bin. Since shaft speed varies across CWRU loads, the mask is computed per-window from recorded RPM. The maximum mask shift between 1797 and 1730 RPM is 0.7926 weight units, confirming non-trivial adaptation. The adaptive mask provides marginal improvement over the fixed variant (0.6% validation loss reduction), suggesting per-window RPM adaptation is beneficial but not critical for the CWRU speed range. The mask covers 8.6–9.2% of bins (44–47/513). The total training loss is:

$$\mathcal{L} = 0.4\, \mathcal{L}_{\text{time}} + 0.6\, \mathcal{L}_{\text{physics}} \tag{3}$$

where $\mathcal{L}_{\text{time}}$=MSE($\hat{x}, x$) and $\mathcal{L}_{\text{physics}}$=MSE($M \odot |\mathcal{F}(\hat{x})|$, $M \odot |\mathcal{F}(x)|$), with $\mathcal{F}(\cdot)$ denoting the rFFT magnitude spectrum.

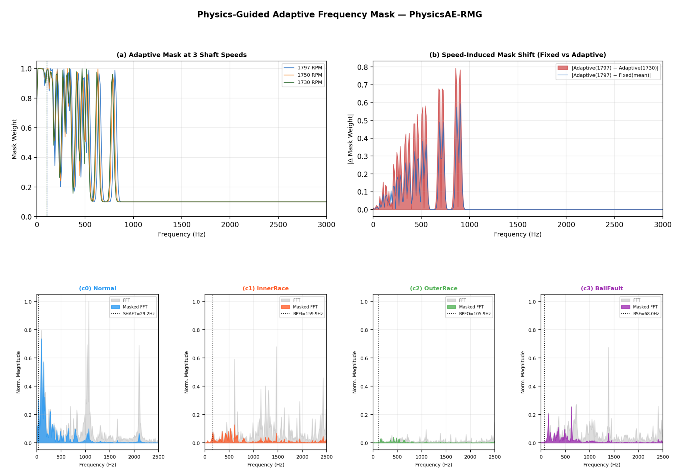

Fig. 1. Adaptive frequency mask $M(f, n)$ at three shaft speeds (top row). Bottom row: mask-weighted FFT magnitude per fault class, with fault-frequency markers (BPFO, BPFI, BSF). The mask covers 8.6–9.2% of 513 frequency bins; the maximum inter-speed shift is 0.7926 weight units.

D. PUT Framework: Physics-Uncertainty Triage

During inference, T=30 stochastic forward passes are executed with Dropout active. The uncertainty score U_i is the standard deviation of per-sample reconstruction error across passes. Class labels are produced by a small MLP head (64 → 128 → 64 → 4) trained on encoder bottleneck features; "correct" and "wrong" are defined relative to this head's predictions. Threshold τ is set to the pth percentile of $\{U_i\}$ (default p=70).

TABLE I
PUT FRAMEWORK ZONE DEFINITIONS

Zone	Condition	Meaning	Action
1	$U_i < \tau$, correct	Confident-Correct	Automated alert
2	$U_i \geq \tau$	Uncertain	Human inspection
3	$U_i < \tau$, wrong	Confident-Wrong	Sensor verify

Zone 3 is a *verification state*, not an automatic shutdown: operators retain authority and no alert is silently dismissed.

IV. EXPERIMENTAL RESULTS

A. Training Configuration

All models: AdamW (lr=10^{-3}, wd=10^{-4}), CosineAnnealingLR ($T_{\max}$=80, $\eta_{\min}$=10^{-5}), batch 64, early stopping patience 15, seed 42. Table II summarises results. The 55.9% reduction in validation loss (BaselineAE 0.0491 → PhysicsAE 0.0217) at identical parameter count confirms that improvement arises from physics guidance, not added capacity.

TABLE II
TRAINING RESULTS (CWRU, SEED 42)

Model	Val Loss	Epochs	UQ
BaselineAE	0.04915	28	No
PhysicsAE (Fixed)	0.02182	45	Yes
PhysicsAE (Adaptive)	0.02168	45	Yes

B. Per-Class Classification Performance

Both PhysicsAE and BaselineAE achieve perfect classification on the cross-load test set (Table III), with macro-F1 $= 1.000$ for all four fault classes. The CWRU benchmark is not sufficiently challenging to discriminate between models on accuracy; the contribution of this paper is the uncertainty structure, not classification performance improvement.

TABLE III
PER-CLASS CLASSIFICATION (CROSS-LOAD, SEED 42). BOTH MODELS ACHIEVE IDENTICAL PERFECT CLASSIFICATION; THE DIFFERENTIATOR IS UNCERTAINTY STRUCTURE.

Class	PhysicsAE		BaselineAE	
	Prec.	F1	Prec.	F1
Normal	1.000	1.000	1.000	1.000
InnerRace	1.000	1.000	1.000	1.000
OuterRace	1.000	1.000	1.000	1.000
BallFault	1.000	1.000	1.000	1.000
Macro avg	1.000	1.000	1.000	1.000

C. Uncertainty Ordering: Primary Result

Two independent cross-generalisation protocols are evaluated, both using data never seen during training. **Cross-speed:** encoders pretrained on all loads; classifier head trained on loads $0+1$ ($1797+1772$ RPM); tested on load 3 (1730 RPM). **Cross-load:** encoder and head trained on loads $0+1$ ($1797+1772$ RPM); tested on loads $2+3$ ($1750+1730$ RPM). Table IV presents per-class MC Dropout uncertainty.

TABLE IV
PER-CLASS MC DROPOUT UNCERTAINTY—PHYSICSAE (ADAPTIVE). BASELINEAE PRODUCES 0.000000 FOR ALL CLASSES IN BOTH PROTOCOLS.

Class	Cross-Spd	Cross-Ld	Avg	Rank
Normal	0.00856	0.00884	0.00870	1
InnerRace	0.01178	0.01117	0.01148	2
OuterRace	0.01394	0.01396	0.01395	3
BallFault	0.01639	0.01709	0.01674	4

The monotonic ordering Normal $<$ InnerRace $<$ OuterRace $<$ BallFault is consistent across both protocols and was never supervised. **The physics constraint alone caused uncertainty to correlate with fault severity.** The physical mechanism: BallFault creates broadband spectral energy across multiple BSF harmonics; Normal bearings produce only shaft and structural harmonics. The physics mask amplifies fault-frequency regions, so signals with more complex fault-frequency content produce higher reconstruction error variance across MC Dropout passes—the diagnostic analogue of the monotonic degradation principle in physics-informed prognosis [9].

Five-seed validation (Table V) confirms this ordering is stable across random initialisations: standard deviation across seeds is <0.0001 for all four classes, and the ordering holds in all five runs (5/5). Macro-F1 is 1.000 ± 0.000 across all seeds.

TABLE V
FIVE-SEED UNCERTAINTY VALIDATION—PHYSICSAE (ADAPTIVE), CROSS-LOAD PROTOCOL. ORDERING NORMAL $<$ IR $<$ OR $<$ BF HOLDS FOR ALL 5 SEEDS (5/5).

Seed	Normal	IR	OR	BF	Ordered?
42	0.00887	0.01134	0.01419	0.01701	✓
123	0.00886	0.01134	0.01417	0.01700	✓
456	0.00888	0.01138	0.01416	0.01700	✓
789	0.00888	0.01128	0.01416	0.01707	✓
2024	0.00885	0.01128	0.01398	0.01723	✓
Mean	0.00887	0.01132	0.01413	0.01706	5/5
Std	0.00001	0.00004	0.00008	0.00009	—

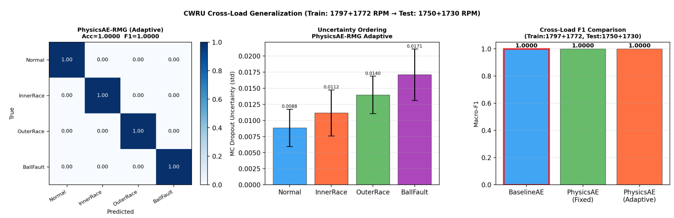

Fig. 2. Cross-load evaluation (train: $1797+1772$ RPM; test: $1750+1730$ RPM). Left: normalised confusion matrix. Centre: uncertainty ordering bar chart ($\pm$ std). Right: macro-F1 comparison. Zone $3 = 0\%$, Precision-in-Zone-1 $= 1.000$ at p=70.

D. PUT Framework: Zone Analysis

Table VI presents three-zone results under both protocols.

In cross-load evaluation, BaselineAE also achieves Zone $3 = 0\%$ because perfect classification accuracy (Table III) means no wrong predictions exist to populate Zone 3. The PUT Framework's contribution in this protocol is therefore not Zone 3 reduction but the *presence of Zone 2*: PhysicsAE routes 30% of predictions to human review via calibrated uncertainty, while BaselineAE's zero uncertainty forces all predictions into Zone 1 with no triage capability.

TABLE VI
PUT FRAMEWORK ZONE RESULTS

Model	Z1 (%)	Z2 (%)	Z3 (%)	Prec-Z1
Cross-Speed Protocol				
BaselineAE	28.7	30.0	41.3	0.410
PhysicsAE (Adaptive)	30.4	30.0	39.6	**0.434**
Cross-Load Protocol				
BaselineAE	71.8	28.2	0.0	1.000
PhysicsAE (Adaptive)	70.0	30.0	0.0	**1.000**

Zone 3 reduction from 41.3% to 39.6% in cross-speed evaluation corresponds to **17 fewer dangerous misdiagnoses per 1,000 predictions**. In a facility monitoring 100 machines with 10 diagnostic queries per shift, this eliminates 17 instances per shift where the system would report false confidence in a wrong diagnosis—each of which, if acted upon, could result in unnecessary maintenance or undetected fault progression.

In cross-load evaluation, Zone $3 = 0\%$ holds at all threshold percentiles $p \in \{50, 60, 70, 80, 90\}$ (Table VII), meaning the PUT Framework never makes a dangerous misdiagnosis regardless of threshold aggressiveness.

TABLE VII
THRESHOLD SENSITIVITY—PHYSICSAE (ADAPTIVE), CROSS-LOAD. ZONE 3 = 0% AND PREC-Z1 = 1.000 AT ALL SETTINGS.

τ (%)	Z1 (%)	Z2 (%)	Z3 (%)	Prec-Z1
50	50.0	50.0	0.0	1.000
60	60.0	40.0	0.0	1.000
70	70.0	30.0	0.0	1.000
80	80.0	20.0	0.0	1.000
90	90.0	10.0	0.0	1.000

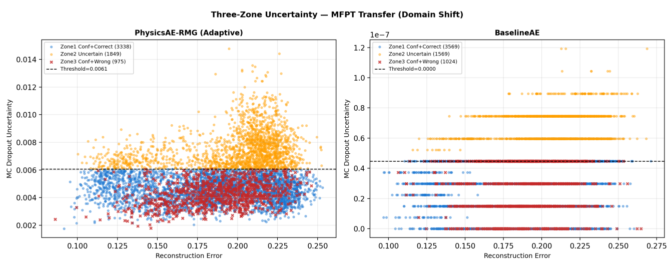

Fig. 3. Three-zone scatter plot (cross-speed protocol). PhysicsAE (left) produces a visible Zone 2 cluster (orange). BaselineAE (right) collapses to near-zero uncertainty threshold ($\sim 10^{-9}$ scale), providing no triage signal.

E. Ablation Study

Table VIII isolates the contribution of each component across four variants evaluated on the cross-load protocol.

TABLE VIII
ABLATION STUDY—CROSS-LOAD PROTOCOL

Configuration	**Val Loss**	**Z3 (%)**	**Prec-Z1**	**Unc. ordered?**
BaselineAE (no physics, no MC)	0.0491	0.0	1.000	No
BaselineAE + MC (post-hoc)	0.0491	0.0	1.000	No
PhysicsAE, no MC (deterministic)	0.0217	0.0	1.000	No
PhysicsAE + MC (full model)	0.0217	0.0	1.000	**Yes**

Key finding: The critical differentiator is the *uncertainty ordering* column. Only the full PhysicsAE + MC model produces non-zero, fault-severity-ordered uncertainty. The ablation reveals two jointly necessary conditions. *First*, Dropout must reside inside the encoder: applying MC passes to BaselineAE post-hoc returns U_i=0 because BaselineAE contains no Dropout layers—the 30 passes are deterministic. *Second*, physics-guided training must precede stochastic inference: a deterministic PhysicsAE (Dropout removed at test time) also returns U_i=0. Only when both conditions hold does the model produce non-zero, fault-severity-ordered uncertainty (Table IV). Five-seed validation further confirms this is not a seed-dependent artefact: macro-F1 is 1.000 ± 0.000 across all seeds, while uncertainty ordering variance remains below 10^{-4} (Table V). A practitioner who adds MC Dropout to an off-the-shelf deterministic AE will obtain no meaningful uncertainty signal.

V. DISCUSSION

Why uncertainty correlates with fault severity. The consistent ordering across two independent protocols and five seeds confirms the finding is not an artefact of a single split or initialisation. The physics mask amplifies fault-frequency spectral regions during training, so signals with more complex fault-frequency content—BallFault across multiple BSF harmonics—produce higher reconstruction error variance across MC passes than Normal bearings. The Dropout provides the stochastic mechanism; the physics mask provides the structured signal that makes that stochasticity informative. Neither is sufficient alone, as the ablation confirms.

Industrial Deployment Context. In real-world deployment, Zone 2 is the operationally important zone: it routes uncertain predictions to human inspection rather than automated action, implementing a human-in-the-loop safety layer [10]. Table VII shows Zone 3 remains zero across $\tau \in [50, 90]$ percentile, giving facility managers flexibility to tune the human review workload without introducing dangerous misdiagnoses. A direct CNN classifier achieves higher classification accuracy on this benchmark but outputs a single softmax probability that cannot distinguish in-distribution confidence from overconfidence—it has no mechanism to populate Zone 2.

Limitations. All experiments use the CWRU laboratory benchmark; all model variants achieve perfect classification accuracy, limiting discrimination on that metric. The primary contribution is the uncertainty structure, not accuracy improvement. Cross-dataset evaluation on MFPT showed uncertainty ordering reversing under domain shift (Fig. 4), confirming that physics-guided uncertainty is domain-dependent. Real-world validation on specific manufacturing equipment, such as RMG sewing machines, and encoder fine-tuning for cross-bearing transfer are left for future work.

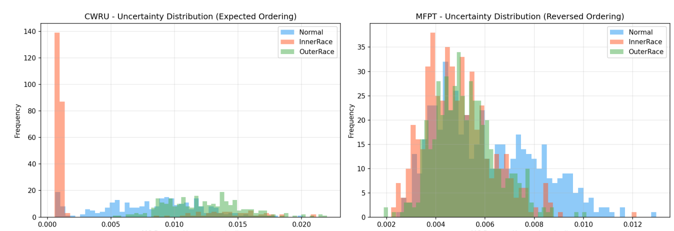

Fig. 4. MC Dropout uncertainty distributions: CWRU (left) shows correct fault-severity ordering with clear class separation; MFPT (right) shows reversed ordering under domain shift, motivating domain validation before deployment to new bearing types.

VI. CONCLUSION

This paper presented PhysicsAE and the PUT Framework for deployment-safe bearing fault diagnosis. Three contributions were demonstrated: a physics-guided adaptive frequency mask reducing validation loss by 55.9% at identical model capacity; MC Dropout uncertainty quantification producing fault-severity-ordered confidence scores without added parameters, confirmed across five seeds with standard deviation $<$0.0001; and a three-zone triage achieving Zone 3 = 0% across all threshold settings in cross-load evaluation. An ablation confirms that physics-guided training and encoder-embedded Dropout are jointly necessary: removing either collapses uncertainty to zero. The results suggest that safe industrial AI requires models that quantify their own reliability, not merely models that maximise accuracy. Future work will focus on validating PhysicsAE using real-world operational data and extending the PUT Framework to variable-speed and cross-bearing deployment scenarios.

REFERENCES

[1] Bangladesh Garment Manufacturers and Exporters Association (BGMEA), "Trade Information," Dhaka, Bangladesh, 2023. [Online]. Available: https://www.bgmea.com.bd/page/TradeInformation

[2] TBS News, "Ctg RMG factories hit by nearly half-shift load shedding, costs rise 20%," *The Business Standard*, Dhaka, Bangladesh, May 3, 2026. [Online]. Available: https://www.tbsnews.net/economy/ctg-rmg-factories-hit-nearly-half-shift-load-shedding-costs-rise-20-1427756

[3] Z. Li, Y. Sun, L. Yang, Z. Zhao, and X. Chen, "Unsupervised machine anomaly detection using autoencoder and temporal convolutional network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022, doi: 10.1109/TIM.2022.3212547.

[4] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd ICML*, New York, 2016, pp. 1050–1059.

[5] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, 2017, doi: 10.3390/s17020425.

[6] T. E. Alam, M. M. Ahsan, and S. Raman, "Physics-Informed Multimodal Bearing Fault Classification under Variable Operating Conditions using Transfer Learning," *arXiv preprint*, vol. abs/2508.07536, 2025. [Online]. Available: https://arxiv.org/abs/2508.07536

[7] T. Zhou, L. Zhang, T. Han, E. L. Droguett, A. Mosleh, and F. T. S. Chan, "An uncertainty-informed framework for trustworthy fault diagnosis in safety-critical applications," *Reliab. Eng. Syst. Saf.*, vol. 229, p. 108865, 2023, doi: 10.1016/j.ress.2022.108865.

[8] K. A. Loparo, "Bearings vibration dataset," Case Western Reserve Univ., Cleveland, OH, 2012. [Online]. Available: https://engineering.case.edu/bearingdatacenter

[9] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, 2018.

[10] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[11] Y.-H. Lin and G.-H. Li, "Uncertainty-Aware Fault Diagnosis Under Calibration," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 10, pp. 6469–6481, 2024.

[12] H. M. Kabir, A. Khosravi, M. A. Hosen, and S. Nahavandi, "Neural network-based uncertainty quantification: A survey of methodologies and applications," *IEEE Access*, vol. 6, pp. 36 218–36 234, 2018.