MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Interpretable Physics-Informed Multimodal Deep Learning for Eccentricity-Severity Estimation of Induction Machines

Su, Hanqi; Liu, Dehong; Inoue, Hiroshi; Wang, Yebin

TR2026-121 September 09, 2026

Abstract

The problem of estimating eccentricity severity in induction machines under different and unseen conditions is addressed in this paper. It is challenging to achieve accurate and reliable estimations for fault severity. Unimodal approaches often fail to capture complementary information across heterogeneous sensors to achieve accurate results, while deep-learning models frequently lack interpretability and tend to overfit when the dataset is small, resulting in unstable results, especially when the motor is operating at varying conditions. To overcome these issues, we propose a physics-informed multimodal deep learning framework that integrates physics-informed and statistical features extracted from current, vibration, and air-gap signals. We first design a unimodal regression backbone with learnable scaling and decorrelation penalties to extract compact latent representations. These representations are then fused using a cross-modal attention mechanism, enabling adaptive information sharing across modalities. Model interpretability is enhanced using SHapley Additive exPlanations (SHAP) to identify physically meaningful features. Experimental results demonstrate that our framework achieves the best or second-best performance across different test settings, with explanatory power improvements of 4.4%–10% over unimodal baselines and greater stability compared to conventional fusion strategies.

International Conference on Electrical Machines (ICEM) 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Interpretable Physics-Informed Multimodal Deep Learning for Eccentricity-Severity Estimation of Induction Machines

Hanqi Su
University of Maryland, College Park
College Park, MD, USA
hanqisu@umd.edu

Dehong Liu
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA, USA
liudh@merl.com

Hiroshi Inoue
Mitsubishi Electric Corporation
Amagasaki, Hyogo, Japan
Inoue.Hiroshi@cw.MitsubishiElectric.co.jp

Yebin Wang
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA, USA
yebinwang@merl.com

***Abstract*—The problem of estimating eccentricity severity in induction machines under different and unseen conditions is addressed in this paper. It is challenging to achieve accurate and reliable estimations for fault severity. Unimodal approaches often fail to capture complementary information across heterogeneous sensors to achieve accurate results, while deep-learning models frequently lack interpretability and tend to overfit when the dataset is small, resulting in unstable results, especially when the motor is operating at varying conditions. To overcome these issues, we propose a physics-informed multimodal deep learning framework that integrates physics-informed and statistical features extracted from current, vibration, and air-gap signals. We first design a unimodal regression backbone with learnable scaling and decorrelation penalties to extract compact latent representations. These representations are then fused using a cross-modal attention mechanism, enabling adaptive information sharing across modalities. Model interpretability is enhanced using SHapley Additive exPlanations (SHAP) to identify physically meaningful features. Experimental results demonstrate that our framework achieves the best or second-best performance across different test settings, with explanatory power improvements of 4.4%–10% over unimodal baselines and greater stability compared to conventional fusion strategies.**

***Index Terms*—Multimodal learning, Physics-informed learning, Fault detection, Model interpretability**

I. Introduction

Eccentricity is one of the most common faults of induction machines [1], [2], where the rotor axis is misaligned with the stator axis, leading to an unbalanced air gap. In static eccentricity, the position of the minimum air-gap length is fixed, whereas in dynamic eccentricity it rotates with the rotor. Studies have shown that both types of eccentricity often coexist, as an inherent level of static eccentricity is present even in newly manufactured machines due to assembly tolerances, and prolonged operation can further cause bearing wear, rotor-shaft bending, and other mechanical issues [3], [4], leading to dynamic eccentricity. Such eccentricity faults degrade motor performance by inducing torque ripple, unbalanced magnetic pull, and abnormal vibration, and in severe cases may result in insulation damage or sudden breakdown [5]. Therefore, accurate detection and severity estimation of eccentricity are critical for preventive maintenance in industrial applications [6]–[8].

Traditional diagnostic approaches rely on physics-based analysis of individual sensing modalities such as current or vibration. Motor current signature analysis (MCSA) is widely used, as eccentricity often produces characteristic sideband components in the current spectrum [9], [10]. Vibration-based methods, often combined with statistical measures such as kurtosis or root mean square, are also effective for capturing mechanical imbalance [11], [12]. Air-gap sensors provide complementary measurements reflecting geometric deviations directly [13], [14]. While these unimodal methods are interpretable and grounded in physical knowledge, they are inherently limited in robustness. A single modality may fail under certain load or fault conditions because unimodal models cannot fully capture the complementary information available across heterogeneous sensing channels. For example, excessive vibration may be observed in the early stages of a mechanical fault, whereas the current often exhibits few or no discernible fault signatures.

Recent advances in machine learning have motivated the application of deep neural networks for intelligent fault diagnosis [2], [15], [16]. Convolutional [17], [18], recurrent [19], and transformer-based architectures [20] have demonstrated strong predictive performance when applied to time-series and spectral representations. In the meantime, with the growing availability of multimodal sensing data, multimodal learning has the potential to improve robustness and generalization beyond what unimodal models can achieve [21]–[23]. Existing fusion strategies generally fall into three categories: early

Hanqi Su performed this work as an intern at MERL.

fusion, late fusion, and intermediate fusion [23]. Although these methods typically outperform unimodal baselines, they remain static in their integration strategies and fail to adaptively capture modality-specific dependencies. At a broader level, recent work in industrial artificial intelligence (IAI) has highlighted broader system-level requirements. Frameworks such as the industrial large-knowledge model [24] and unified foundations for industrial AI [25] emphasize that effective AI solutions must not only achieve high accuracy, but also integrate domain knowledge, enable multimodal information fusion, and provide interpretable insights. However, existing methods for eccentricity diagnosis rarely satisfy these combined requirements, leaving a gap between industrial AI visions and practical implementations.

Despite these advances, several key challenges remain. First, deep models often operate as black boxes, providing little insight into which features drive their predictions [26], [27]. This lack of interpretability restricts trust and adoption in industrial settings, where engineers often require explanations aligned with physical principles. Second, industrial datasets are typically limited in size and coverage, with imbalanced or missing fault conditions. This data scarcity poses difficulties for large data-hungry models and calls for designs that incorporate physics-informed priors and regularization.

Motivated by these challenges, this work proposes an interpretable physics-informed multimodal deep learning framework for eccentricity severity estimation of induction machines. The core contributions are as follows:

- We propose a physics-informed and statistical feature extraction scheme for multimodal signals, ensuring physically meaningful representations.
- We design a novel unimodal regression backbone with learnable scaling and decorrelation penalties to produce compact and informative latent features and then develop a cross-modal attention mechanism for adaptive multimodal fusion that outperforms baseline fusion methods.
- We enhance model interpretability through Shapley Additive exPlanations (SHAP) analysis, which highlights feature importance and validates that the model emphasizes physically consistent indicators.

II. Proposed Method

As illustrated in Fig. 1, we propose an end-to-end multimodal deep learning framework that integrates physics-informed and statistical features with a cross-modal attention-based deep neural architecture, further enhanced by SHAP analysis for model interpretability. This framework utilizes heterogeneous sensing modalities—specifically vibration, current, and air-gap measurements—collected from an industrial motor system. The detailed methodology is presented in the following subsections.

A. Physics-Informed and Statistical Feature Extraction

To ensure comprehensive representation across sensing modalities, we extract both physics-informed and statistical features from acceleration, current, and air-gap signals in both the time and the frequency domains.

For physics-informed features, physically interpretable indicators related to the power quality and operating conditions are extracted. In the time domain, we compute the root mean square (RMS), peak-to-peak value, and crest factor to capture variations in signal amplitudes, impulsive behaviors, and transient responses associated with eccentricity faults. In the frequency domain, we extract energy features over multiple frequency bands (0–500Hz, 500–2000Hz, 2000-5000Hz), respectively, dominant frequency component magnitudes (0–500Hz, 500–2000Hz), and total harmonic distortion (THD), etc. In addition, sideband features are computed around the supply frequency f_s. These sidebands are used in motor current signature analysis (MCSA) to detect eccentricity faults through patterns at $f_s \pm k f_r$, where f_r denotes the rotational frequency and k is an integer. Although the sideband interpretation is less direct for vibration and air-gap signals, modulation effects arising from mechanical asymmetry may still manifest as spectral patterns, making such features useful when applied across modalities.

To complement these physics-based features, statistical features are also extracted, including mean, absolute mean, standard deviation, skewness, and kurtosis in the time domain; and frequency mean, variance, spectral entropy, spectral kurtosis, and skewness in the frequency domain. Statistical features are especially important for air-gap signals, as they capture variations in rotor–stator alignment and magnetic field distribution beyond what sideband analysis alone can provide. Physics-informed and statistical features are then combined to form a unified representation, providing a compact and informative basis for regression modeling.

B. Novel Unimodal Regression Model and Multimodal Attention Fusion

1) Novel unimodal regression model development: Before constructing the multimodal architecture, it is essential to develop robust unimodal regression models that can effectively capture the useful patterns within each sensing modality. A naive multilayer perceptron (MLP) regression model may overfit noisy features or fail to differentiate between redundant representations, which would propagate errors into the multimodal stage. To address this, we propose a novel unimodal regression design that introduces feature-wise learnable scaling, a compact bottleneck representation, and dedicated regularization terms for sparsity and decorrelation, as shown in Fig. 1.

Given an input feature vector $\mathbf{x}_m \in \mathbb{R}^{d_m}$ from the m^{th} modality, the first step is normalization followed by element-wise scaling with a learnable vector $\mathbf{s}_m \in \mathbb{R}^{d_m}$

$$\tilde{\mathbf{x}}_m = \mathbf{s}_m \odot \mathrm{Norm}(\mathbf{x}_m), \tag{1}$$

where $\odot$ denotes element-wise multiplication. This scaling mechanism adaptively controls the contribution of each feature, suppressing irrelevant signals while emphasizing infor-

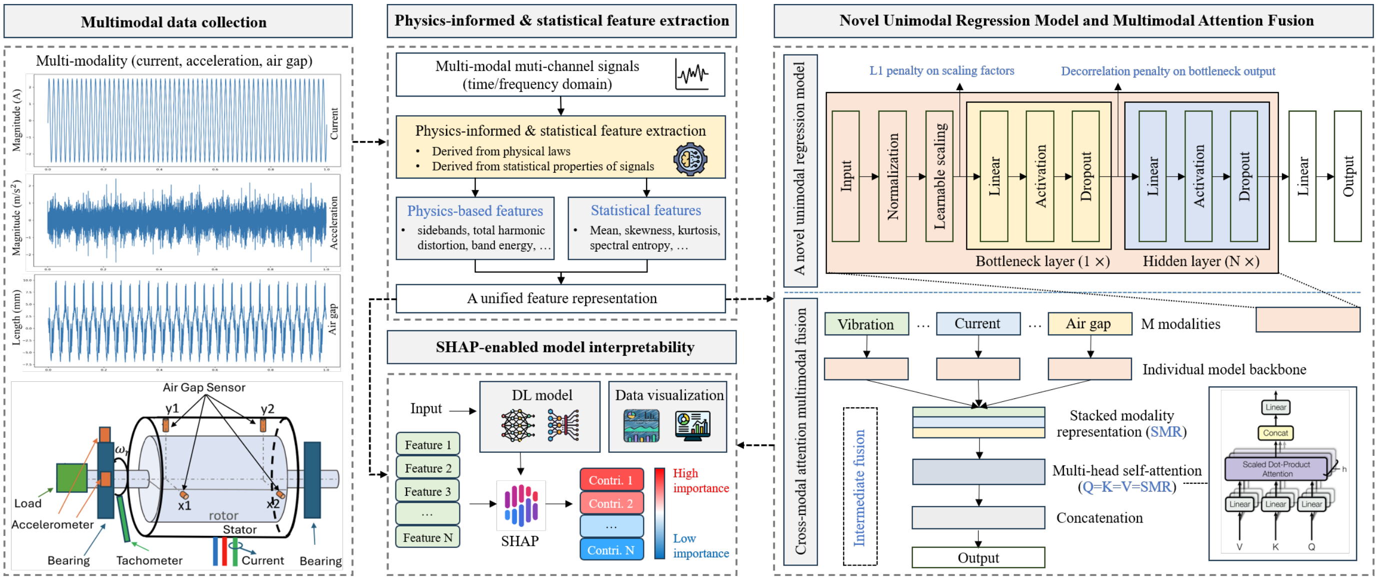

Fig. 1. Proposed end-to-end multimodal deep learning framework.

mative components. To prevent over-reliance on noisy features, the scaling vector is regularized by an L_1-penalty,

$$\mathcal{L}_{\text{scale}} = \alpha \|\mathbf{s}_m\|_1, \tag{2}$$

where $\alpha > 0$ is a regularization coefficient.

The bottleneck layer serves to compress the scaled features into a lower-dimensional space, which not only improves generalization but also forces the network to extract compact and informative representations. To achieve this, the scaled features are then mapped into a compact bottleneck representation through a linear transformation, non-linear activation, and dropout

$$\mathbf{z}_m = \text{Dropout}(\sigma(W_b \tilde{\mathbf{x}}_m + \mathbf{b}_b)), \tag{3}$$

where $W_b \in \mathbb{R}^{d_b \times d_m}$ and $\mathbf{b}_b \in \mathbb{R}^{d_b}$ denote the weight matrix and bias vector of the bottleneck layer, respectively, d_b is the bottleneck dimension, and $\sigma(\cdot)$ represents a non-linear activation function. To reduce redundancy among bottleneck dimensions, we introduce a decorrelation penalty

$$\mathcal{L}_{\text{decor}} = \beta \sum_{i \neq j} \text{Corr}(\mathbf{z}_{m,i}, \mathbf{z}_{m,j})^2, \tag{4}$$

where $\beta > 0$ is a regularization coefficient and $\text{Corr}(\cdot, \cdot)$ denotes the Pearson correlation coefficient between dimensions i and j. This step ensures that different neurons in the bottleneck capture complementary aspects of the signal rather than overlapping information. The bottleneck output $\mathbf{z}_m$ is further processed by N hidden layers, each consisting of a linear layer, non-linear activation, and dropout, producing the latent representation $\mathbf{h}_m \in \mathbb{R}^{d_h}$.

The unimodal prediction is then obtained as

$$\hat{y}_m = W_m \mathbf{h}_m + b_m, \tag{5}$$

where $W_m \in \mathbb{R}^{1 \times d_h}$ and $b_m \in \mathbb{R}$ are the parameters of the regression head, and d_h is the hidden dimension. Finally, the unimodal training objective combines the regression error with the two regularization terms

$$\mathcal{L}_m = \mathcal{L}_{\text{MSE}}(\hat{y}_m, y) + \mathcal{L}_{\text{scale}} + \mathcal{L}_{\text{decor}}, \tag{6}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean squared error between the predicted output $\hat{y}_m$ and the ground truth y. Through this design, each modality is able to produce a sparse, decorrelated, but information-rich representation. This establishes a reliable foundation for subsequent multimodal attention fusion, ensuring that cross-modal learning builds on well-structured unimodal models.

2) Cross-modal attention multimodal fusion: While unimodal models capture modality-specific characteristics, they are inherently limited in exploiting the complementary information available across heterogeneous sensors. In practice, different modalities may exhibit partial overlap while also capturing distinct information, and analyzing them jointly provides more reliable and robust predictions than relying on any single modality alone. To achieve such integration, we adopt a cross-modal attention mechanism that enables each modality to selectively attend to informative patterns in others [28]. In our design, we adopt the unimodal regression model structure up to its last hidden layer as the backbone to extract latent representations for each modality. These representations are then combined to form a stacked modality representation (SMR). Specifically, if $\mathbf{h}_m \in \mathbb{R}^{d_h}$ denotes the hidden-layer output from modality m, then the stacked representation is

$$\mathbf{H} = \left[\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M\right]^\top \in \mathbb{R}^{M \times d_h}, \tag{7}$$

where M is the total number of modalities and each row of $\mathbf{H}$ corresponds to the hidden representation of one modality. To capture cross-modal dependencies, we adopt a multi-head self-attention mechanism in which the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are all set equal to $\mathbf{H}$ ($\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{H}$) [28].

The scaled dot-product attention is then computed as a fused representation

$$\mathbf{H}_F = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{H}, \tag{8}$$

where d_k is the key dimension. Multiple attention heads are applied in parallel, enabling the model to attend to different modalities simultaneously. By stacking latent representations, self-attention allows each modality to selectively attend to informative patterns from others while preserving its own contextual information. The output of the attention block has the same dimension as $\mathbf{H}$, ensuring compatibility with the unimodal backbones. These attended representations are concatenated and passed through a regression layer to generate the final prediction

$$\hat{y} = \mathbf{W}_F\mathbf{h}_F + b_F, \tag{9}$$

where $\mathbf{h}_\mathrm{F}$ is the concatenated attended representation $\mathbf{H}_F$, $\mathbf{W}_F \in \mathbb{R}^{1\times(M\cdot d_h)}$, and $b_F \in \mathbb{R}$.

The overall multimodal training objective integrates the regression error with the unimodal penalties

$$\mathcal{L} = \mathcal{L}_{\text{MSE}}(\hat{y}, y) + \sum_{m=1}^{M}\left(\mathcal{L}_{\text{scale}}^{(m)} + \mathcal{L}_{\text{decor}}^{(m)}\right), \tag{10}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean squared error between prediction $\hat{y}$ and ground truth y. Through this design, the framework leverages cross-modal attention to exploit complementary sensing information.

C. SHAP-Enabled Model Interpretability

Deep multimodal regression models often operate as black boxes, making it difficult to determine which features or modalities drive the predictions. To improve transparency, we integrate SHapley Additive exPlanations (SHAP) [29] into the proposed framework. SHAP is a model-agnostic interpretability method that assigns an importance value to each input feature by estimating its marginal contribution to the model output, based on principles of cooperative game theory. Formally, given an input feature vector $\mathbf{x} \in \mathbb{R}^d$ and a trained model $f(\mathbf{x})$, the prediction can be expressed as a linear decomposition of feature contributions

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{d}\phi_i, \tag{11}$$

where ϕ_0 is the baseline output (expected value of the model prediction across a background distribution), and ϕ_i is the SHAP value corresponding to the i-th feature. Each ϕ_i represents the marginal contribution of that feature to the prediction relative to ϕ_0.

By applying SHAP to our multimodal regression framework, we obtain feature-level attribution scores across all modalities. This enables us to quantify the relative importance of both physics-informed and statistical features. Visualizations of SHAP values provide interpretable explanations of the model's behavior, allowing domain experts to verify whether the model emphasizes physically meaningful features. In this way, SHAP bridges the gap between high-performance regression models and interpretable decision support.

III. Experiments

A. Experimental setup

Dataset. The experimental platform is based on a 0.75 kW three-phase squirrel-cage induction motor, as illustrated in Fig. 1. To generate controlled static eccentricity, the two original bearings between the rotor and stator are replaced by larger external bearings, which allow manual adjustment of the rotor position with high accuracy. A magnetic powder brake is employed as the load, whose torque is tuned by changing the input operating current. The entire system is enclosed in a protective clear cage for safety. During operation, multiple synchronized signals are acquired from heterogeneous sensors: three-phase stator currents (i_u, i_v, i_w), two orthogonal accelerometers (horizontal and vertical vibration v_x, v_y), and four air-gap sensors $(g_{x1}, g_{y1}, g_{x2}, g_{y2})$. Experiments are conducted across five normalized eccentricity levels $y_i \in \{0, 0.11, 0.25, 0.43, 0.56\}$, where 0 indicates healthy condition and 1 indicates the worst possible faulty condition. For each eccentricity level, eight torque loads $T_j \in \{0.0, 0.3, 0.5, 0.9, 1.4, 2.0, 2.7, 3.5\}$Nm are considered in experiments, resulting in a total of 40 experiments, each with multimodal time-series data collected for 60 seconds for further analysis.

Train–Test Split. To better reflect realistic application scenarios, our splitting strategy is designed such that the testing stage may involve unseen conditions, including both load settings not present in the training data and eccentricity levels absent from training. This setting allows us to evaluate both robustness and interpolation capability.

The following train-test split process is performed across all modalities. We take 100% sequences of one eccentricity level (0.11, 0.25, or 0.43) and 25% sequences of each of the remaining four levels as the testing dataset. All remaining sequences are used as the training dataset.

Each original sequence is segmented into 60 non-overlapping 1-second samples, yielding 480 samples per eccentricity level. Based on our train-test split strategy, a total of 960 samples are used for testing, where 480 samples are from one severity level, and the remaining 480 samples are from the remaining four severity levels, with 120 each level. A total of 1440 samples are used for training, in which 80% samples are used for training, and 20% are reserved for validation. This design ensures a fair comparison between models while enabling systematic evaluation of generalization to unseen labels and load conditions.

Implementation Details. We extract both physics-informed and statistical features across modalities. In total, we generate 172 features (67 sideband frequency components for $k = -1, 0, ..., 65$ and 19 other features for x and y channels) for acceleration modality, 124 features (67 sideband frequency components for $k = -1, 0, ..., 65$, for one aggregated channel and 19 other features for three-phase channels) for current

TABLE I
PERFORMANCE COMPARISON OF UNIMODAL REGRESSION MODELS AND MULTIMODAL FUSION STRATEGIES UNDER THREE SETTINGS (HOLDOUT DATA OF SEVERITY LEVELS 0.11, 0.25, AND 0.43, RESPECTIVELY).

Settings			**Severity level 0.11**		**Severity level 0.25**		**Severity level 0.43**	
Model		Modality	**MSE** ($\times 10^{-3}$)	R^2	**MSE** ($\times 10^{-3}$)	R^2	**MSE** ($\times 10^{-3}$)	R^2
Unimodal		Vibration	4.57 ± 2.52	0.857 ± 0.079	9.75 ± 0.83	0.628 ± 0.031	9.88 ± 0.74	0.691 ± 0.023
		Current	4.24 ± 0.52	0.867 ± 0.017	2.52 ± 0.70	0.904 ± 0.027	5.95 ± 2.96	0.815 ± 0.092
		Air gap	3.57 ± 0.77	0.889 ± 0.024	5.97 ± 1.36	0.772 ± 0.052	1.97 ± 0.91	0.939 ± 0.029
Multimodal	Early fusion	Vibration, Current, & Air gap	1.94 ± 1.32	0.940 ± 0.041	3.56 ± 1.23	0.864 ± 0.047	1.87 ± 0.70	0.942 ± 0.022
	Late fusion		0.72 ± 0.33	0.978 ± 0.010	1.93 ± 0.67	0.926 ± 0.026	0.98 ± 0.79	0.970 ± 0.025
	Intermediate summation		0.54 ± 0.14	0.983 ± 0.004	1.63 ± 0.62	0.938 ± 0.024	0.88 ± 0.80	0.972 ± 0.025
	Intermediate concatenation		**0.44 ± 0.21**	**0.986 ± 0.006**	1.49 ± 0.77	0.943 ± 0.030	0.77 ± 0.52	0.976 ± 0.016
	Cross-modal attention		0.64 ± 0.13	0.980 ± 0.004	**1.36 ± 0.58**	**0.948 ± 0.022**	**0.54 ± 0.15**	**0.983 ± 0.005**

modality, and 108 features (16 sideband frequency components for $k = -1, 0, ..., 14$ for both x and y channels and 19 other features for all four x1, x2, y1, and y2 channels) for air-gap modality, where the sideband order k is determined based on correlation analysis and exploratory data analysis. For the unimodal regression model, the bottleneck and hidden layer dimensions are both set to 64, with LeakyReLU (slope 0.1) activations [30]. Dropout rates are set to 0.1 for the bottleneck and 0.3 for hidden layers. Regularization includes an L_1 penalty ($\alpha = 10^{-5}$) and a decorrelation penalty ($\beta = 0.001$). These values are selected empirically to balance regularization and training stability. For example, it is observed that larger dropout values may cause underfitting and unstable convergence in preliminary experiments. For the multimodal cross-attention model, the model parameter settings are the same, and the number of attention heads is fixed to 4. Optimization is performed using Adam with different learning rates for scaling and backbone parameters [31]. For unimodal training, the scaling layer uses a learning rate of 0.005, while the bottleneck and hidden layers use 0.001. For multimodal training, the scaling layers use 0.0015, and the other layers use 0.0003. A weight decay of 10^{-5} is applied in all cases. Models are trained with a batch size of 128 for up to 300 epochs, using early stopping with a patience of 15 epochs.

All experiments are repeated 10 times, and we report the mean and standard deviation. Models are implemented in PyTorch. Performance is evaluated using mean squared error (MSE) and coefficient of determination (R^2)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{13}$$

where y_i represents the ground-truth severity of the ith sample, $\hat{y}_i$ represents the estimated value, and $\bar{y}$ represents the mean of ground-truth values of all samples.

B. Experimental results

We first evaluate the performance of unimodal regression models trained separately on vibration, current, and air-gap features. To validate the effectiveness of our proposed framework, we benchmark it against several standard multimodal fusion strategies: (i) early fusion, where features are concatenated before regression; (ii) late fusion, where unimodal predictions are combined via learnable weights; and (iii) intermediate fusion, where latent features are either combined through a learnable weighted summation across modalities or concatenated directly.

As shown in Table I, both unimodal and multimodal models are evaluated under three settings, where eccentricity levels 0.11, 0.25, and 0.43 are fully held out for testing, respectively. For unimodal regression, the best-performing modality varies across cases. When severity levels 0.11 and 0.43 are fully held out respectively, the air-gap model achieves the lowest mean squared error (MSE) and highest R^2, indicating that this modality is particularly sensitive to eccentricity under different operating conditions. In contrast, when severity level 0.25 is held out, the current-based model provides the best performance, while vibration consistently lags behind the other two modalities across all settings. This variation highlights the complementary nature of the modalities and the importance of combining them for robust regression.

Multimodal fusion approaches—including early, late, and intermediate fusion—all outperform unimodal regression models (except early fusion in holdout 0.25 case). These multimodal models consistently achieve lower MSE and higher R^2, demonstrating that integrating heterogeneous signals improves predictive performance. On average, the explanatory power represented by R^2 of multimodal models is 4.4%–10% higher than that of the best unimodal models, confirming the benefit of cross-modality information sharing. Among the fusion strategies, cross-modal attention achieves the best or the second-best results in most cases. When data of severity level 0.11 are held out for testing, it yields an R^2 of 0.980, slightly lower than intermediate concatenation (0.986). When data of severity level 0.25 are held out for testing, it achieves an R^2 of 0.948, outperforming all other fusion methods. While data of severity level 0.43 are held out for testing, it delivers the overall best performance with an R^2 of 0.983 and an MSE of only 0.54×10^{-3}. In addition to achieving high R^2 values, cross-modal attention also demonstrates superior stability, with consistently lower variance across MSE and R^2 compared to other methods. This indicates that the proposed fusion strategy provides more consistent predictions under different separation settings. Such stability is particularly important in industrial applications with varying operating conditions, where different sensing modalities may respond differently to fault severity.

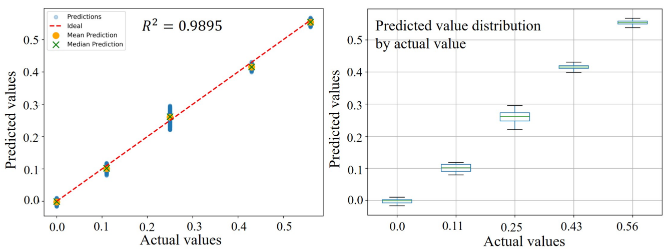

Fig. 2. Estimated versus actual eccentricity severity under the holdout 0.25 setting using cross-modal attention multimodal fusion model, showing (a) regression fit with $R^2 = 0.9895$ and (b) consistent prediction distributions.

To further examine the explanatory capacity of the proposed framework, we analyze the best-performing cross-modal attention model under the holdout 0.25 setting. Fig. 2 illustrates the predicted versus actual values, where the regression fit achieves an R^2 of 0.9895, confirming the strong predictive accuracy and reliability of the model. The distribution plot further shows that predictions across all severity levels are well aligned with the ground truth, indicating robust interpolation capability.

TABLE II
TOP 10 SHAP FEATURE IMPORTANCE RESULTS (TEST DATA) (HOLDOUT SEVERITY LEVEL 0.25)

Ranking	Feature	SHAP feature importance	Modality (channel)
1	Sideband k_{-1} (30Hz)	0.0207	Current (i_u, i_v, i_w)
2	Standard deviation	0.0144	Acceleration (v_x)
3	Sideband k_1 (90Hz)	0.0140	Current (i_u, i_v, i_w)
4	Mean absolute value	0.0137	Air gap (g_{x1})
5	Skewness	0.0132	Current (i_w)
6	Sideband k_2 (120Hz)	0.0095	Air gap (g_{x1}, g_{x2})
7	Mean	0.0075	Air gap (g_{y1})
8	Root mean square	0.0068	Current (i_u)
9	Root mean square	0.0066	Air gap (g_{x1})
10	Total harmonic distortion	0.0036	Current (i_u)

To interpret which features drive the estimation, we apply SHAP analysis. Table II reports the top-10 most important features for the holdout 0.25 test set. Notably, the most influential predictors include sideband features at 30Hz and 90Hz of the current, standard deviation of the vibration, mean absolute value of the air gap, as well as skewness and root mean square. These findings are consistent with domain knowledge: sideband frequencies in motor currents are known indicators of eccentricity, vibration statistics capture structural imbalance, and air-gap measurements directly reflect geometric misalignment. Fig. 3 visualizes the trend of the top-ranked features across eccentricity levels. As expected, current sideband amplitudes increase steadily with fault severity, vibration standard deviation grows as eccentricity rises, and air-gap mean values exhibit clear monotonic behavior. These consistent and physically meaningful patterns demonstrate that the proposed framework not only achieves high predictive performance but also learns interpretable feature relationships that align with engineering knowledge. This highlights the advantage of combining multimodal learning with SHAP-enabled interpretability for reliable fault diagnosis and prognosis.

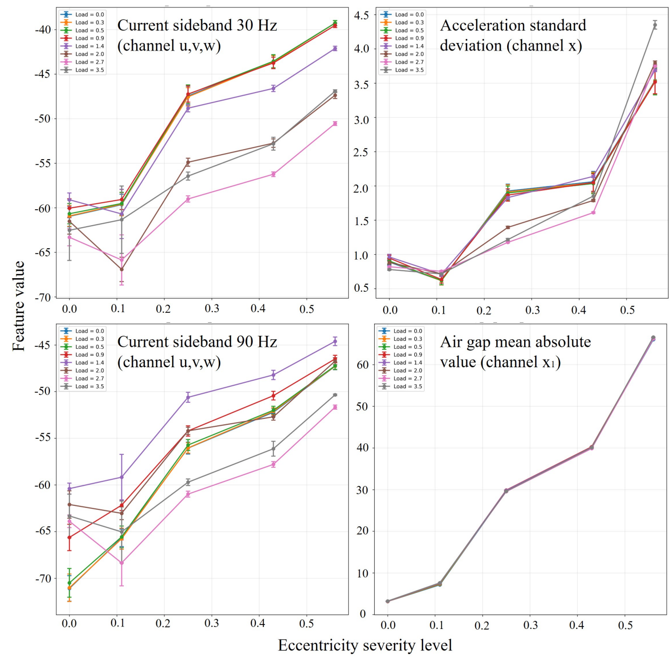

Fig. 3. Top four SHAP-ranked features under the holdout 0.25 setting using cross-modal attention multimodal fusion model.

In addition, the proposed framework has relatively low computational complexity because it operates on compact engineered features and low-dimensional latent representations rather than raw high-dimensional signals. The unimodal regression backbone mainly consists of lightweight fully connected layers, while the attention mechanism is applied only across three modality-level representations instead of long temporal sequences. Compared with end-to-end deep learning-based approaches on raw signals, this design reduces both memory usage and computational cost. Therefore, the proposed framework has the potential for real-time deployment on industrial edge devices or embedded systems.

IV. CONCLUSION

This paper proposes a novel physics-informed multimodal deep learning framework for eccentricity severity estimation. The framework combines physics-informed and statistical features, a novel unimodal regression backbone with learnable scaling and decorrelation penalties, and a cross-modal attention mechanism for adaptive multimodal fusion. To enhance transparency, SHAP analysis is applied, showing that the model emphasizes physically consistent indicators. Experiments on a controlled induction motor testbed demonstrate that our framework consistently outperforms unimodal and conventional fusion baselines with greater stability across different test settings. These findings confirm both the accuracy and interpretability of the proposed approach. Nevertheless, further investigation is still needed before industrial applications, including dynamically varying load conditions, different motor systems, and practical sensor installation variations.

ACKNOWLEDGMENT

The authors would like to thank Bingnan Wang, Jianlin Guo, and Gordon Wichern for their fruitful discussions.

REFERENCES

[1] Jawad Faiz and Hossein Nejadi-Koti, "Eccentricity fault diagnosis indices for permanent magnet machines: state-of-the-art," *IET Electric Power Applications*, vol. 13, no. 9, pp. 1241–1254, 2019.

[2] Zijian Liu, Pinjia Zhang, Shan He, and Jin Huang, "A review of modeling and diagnostic techniques for eccentricity fault in electric machines," *Energies*, vol. 14, no. 14, pp. 4296, 2021.

[3] David G Dorrell, William T Thomson, and Steven Roach, "Analysis of airgap flux, current, and vibration signals as a function of the combination of static and dynamic airgap eccentricity in 3-phase induction motors," *IEEE Transactions on Industry Applications*, vol. 33, no. 1, pp. 24–34, 2002.

[4] Subhasis Nandi, Hamid A Toliyat, and Xiaodong Li, "Condition monitoring and fault diagnosis of electrical motors—a review," *IEEE Transactions on Energy Conversion*, vol. 20, no. 4, pp. 719–729, 2005.

[5] Muhammad Aman Sheikh, Sheikh Tahir Bakhsh, Muhammad Irfan, Nursyarizal bin Mohd Nor, and Grzegorz Nowakowski, "A review to diagnose faults related to three-phase industrial induction motors," *Journal of Failure Analysis and prevention*, vol. 22, no. 4, pp. 1546–1557, 2022.

[6] Yuri Merizalde, Luis Hernández-Callejo, and Oscar Duque-Perez, "State of the art and trends in the monitoring, detection and diagnosis of failures in electric induction motors," *Energies*, vol. 10, no. 7, pp. 1056, 2017.

[7] Hanqi Su and Jay Lee, "Machine learning approaches for diagnostics and prognostics of industrial systems using open source data from phm data challenges: A review," *International Journal of Prognostics and Health Management*, vol. 15, no. 2, 2024.

[8] Jay Lee, Hanqi Su, Dai-Yan Ji, and Takanobu Minami, "Engineering artificial intelligence: framework, challenges, and future direction," *Machine Learning: Engineering*, vol. 1, no. 1, pp. 013001, 2025.

[9] M El Hachemi Benbouzid, "A review of induction motors signature analysis as a medium for faults detection," *IEEE Transactions on Industrial Electronics*, vol. 47, no. 5, pp. 984–993, 2002.

[10] Xiangtian Zheng, Hiroshi Inoue, Makoto Kanemaru, and Dehong Liu, "Eccentricity severity estimation of induction machines using a sparsity-driven regression model," in *2022 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 2022, pp. 1–6.

[11] JR Cameron, WT Thomson, and AB Dow, "Vibration and current monitoring for detecting airgap eccentricity in large induction motors," in *IEE Proceedings B (Electric Power Applications)*. IET, 1986, vol. 133, pp. 155–163.

[12] Noureddine Bessous, SE Zouzou, Salim Sbaa, and W Bentrah, "A comparative study between the mcsa, dwt and the vibration analysis methods to diagnose the dynamic eccentricity fault in induction motors," in *2017 6th International Conference on Systems and Control (ICSC)*. IEEE, 2017, pp. 414–421.

[13] Galina Mirzaeva, Khalid Imtiaz Saad, and Mohsen Ghaffarpour Jahromi, "Comprehensive diagnostics of induction motor faults based on measurement of space and time dependencies of air gap flux," *IEEE Transactions on Industry Applications*, vol. 53, no. 3, pp. 2657–2666, 2016.

[14] Ievgen Zaitsev, Anatolii Levytskyi, Viktoriia Bereznychenko, and Vadim Rassovskyi, "Air-gap sensors for hydro generators and techniques for air-gap eccentricity fault detection and estimation," in *Power Systems Research and Operation: Selected Problems III*, pp. 111–131. Springer, 2023.

[15] Yuanyuan Yang, Md Muhie Menul Haque, Dongling Bai, and Wei Tang, "Fault diagnosis of electric motors using deep learning algorithms and its application: A review," *Energies*, vol. 14, no. 21, pp. 7017, 2021.

[16] Jay Lee, Hanqi Su, Marco Macchi, Adalberto Polenghi, Wei Wu, Zhiheng Zhao, George Q Huang, Kiva Allgood, Devendra Jain, Benedikt Gieger, et al., "2026 roadmap on artificial intelligence and machine learning for smart manufacturing," *Machine Learning: Engineering*, 2026.

[17] Ma'd El-Dalahmeh, Maher Al-Greer, Imran Bashir, Mo'ath El-Dalahmeh, Aykut Demirel, and Ozan Keysan, "Autonomous fault detection and diagnosis for permanent magnet synchronous motors using combined variational mode decomposition, the hilbert-huang transform, and a convolutional neural network," *Computers and Electrical Engineering*, vol. 110, pp. 108894, 2023.

[18] Xiaopeng Liu, Jianfeng Hong, Kang Zhao, Bingxiang Sun, Weige Zhang, and Jiuchun Jiang, "Vibration analysis for fault diagnosis in induction motors using one-dimensional dilated convolutional neural networks," *Machines*, vol. 11, no. 12, pp. 1061, 2023.

[19] Emanuele Principi, Damiano Rossetti, Stefano Squartini, and Francesco Piazza, "Unsupervised electric motor fault detection by using deep autoencoders," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 2, pp. 441–451, 2019.

[20] Farbod Parvin, Jawad Faiz, Yuan Qi, Ahmad Kalhor, and Bilal Akin, "A comprehensive interturn fault severity diagnosis method for permanent magnet synchronous motors based on transformer neural networks," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 11, pp. 10923–10933, 2023.

[21] Dhanesh Ramachandram and Graham W Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[22] Hanqi Su, Binyang Song, and Faez Ahmed, "Multi-modal machine learning for vehicle rating predictions using image, text, and parametric data," in *International design engineering technical conferences and computers and information in engineering conference*. American Society of Mechanical Engineers, 2023, vol. 87295, p. V002T02A089.

[23] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[24] Jay Lee and Hanqi Su, "A unified industrial large knowledge model framework in industry 4.0 and smart manufacturing," *International Journal of AI for Materials and Design*, vol. 1, no. 2, pp. 41–47, 2024.

[25] Jay Lee and Hanqi Su, "Rethinking industrial artificial intelligence: A unified foundation framework," *International Journal of AI for Materials and Design*, vol. 2, no. 2, pp. 56–68, 2025.

[26] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, 2019.

[27] Christoph Molnar, *Interpretable machine learning*, Lulu.com, 2020.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[30] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al., "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*. Atlanta, GA, 2013, vol. 30, p. 3.

[31] Diederik P Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

BIOGRAPHIES

Hanqi Su (S'25) received his B.Eng. degree in Robotics Engineering from Southern University of Science and Technology, Shenzhen, Guangdong, China, in 2023. Previously, he participated in a one-year Special Student Program at the Massachusetts Institute of Technology (MIT) during his fourth undergraduate year. He is currently pursuing his Ph.D. in Mechanical Engineering at University of Maryland, College Park. His research is primarily focused on large language models, prognostics and health management, machine learning, engineering design, and industrial artificial intelligence.

Dehong Liu (S'01-M'03-SM'08) received his B.S., M.S., and Ph.D. degrees all in electrical engineering from Tsinghua University, Beijing, China, in 1997, 1999, and 2002, respectively. From 2003 to 2010, he was with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, first as a Research Associate and then as a Senior Research Scientist. Since 2010, he has been with Mitsubishi Electric Research Laboratories(MERL), Cambridge, MA, USA, where he is now a Lead Research Scientist. His current research interests include electric machine fault diagnosis, electromagnetic compatibility, signal processing, and machine learning, etc.

Hiroshi Inoue received B.E.and M.E. degree from The University of Tokyo, Japan, in 2012 and 2014, respectively. He has been a researcher at Mitsubishi Electric Corporation since April 2014. His primary research focus is motor malfunction diagnosis technology. He is currently a member of the Mitsubishi Electric Advanced Technology R&D Center, Electromechanical Systems Department.

Yebin Wang (S'06-M'10-SM'16) received his Ph.D. in Electrical Engineering from the University of Alberta, Edmonton, Canada, in 2008. Dr. Wang has been with Mitsubishi Electric Research Laboratories in Cambridge, MA, USA, since 2009, and now is a Senior Principal Research Scientist/Senior Team Leader. His research interests include nonlinear control and estimation, optimal control, adaptive and learning systems, and their applications including mechatronic systems. He is the co-author of over 100 papers in the field of Control and System, and the co-inventor of over 40 US patent applications. He was/is an editorial board member of multiple international journals and conferences.