OPEN ACCESS
IOP Publishing | ISECAEP
Journal of Reliability Science and Engineering
J. Reliab. Sci. Eng. **2** (2026) 025302 (23pp)
https://doi.org/10.1088/3050-2454/ae61b4

Reliability assessment of remaining useful life predictions of roller bearings using deep learning models

Santosh Bisoyi[1], Amit Kumar Rathi[1,2,*] and Swarup Mahato[3]

[1] Department of Civil and Infrastructure Engineering, Indian Institute of Technology Jodhpur, Jodhpur, Rajasthan 342030, India
[2] Rishabh Centre Research and Innovation in Clean Energy, Indian Institute of Technology Jodhpur, Jodhpur, Rajasthan 342030, India
[3] Department of Civil Engineering, Indian Institute of Engineering Science and Technology, Shibpur, West Bengal 711103, India

E-mail: akrathi@iitj.ac.in

Received 31 December 2025, revised 23 March 2026
Accepted for publication 19 April 2026
Published 29 April 2026

Abstract

This study focuses on an accurate and reliable prediction of the remaining useful life (RUL) of the bearings. In recent times, machine learning based RUL predictions have gained traction to ensure safety, efficiency, and cost-effectiveness of industrial machinery. However, raw vibration signals acquired from run-to-failure experiments are typically noisy, non-stationary, and high-dimensional, making direct learning of degradation patterns challenging. This often undermines the reliability of the prediction due to inconsistency in the RUL predictions at a given time. This study proposes a reliable multi-stage prognostic framework that integrates robust signal pre-processing, data-driven health-state classification, and deep learning-based RUL estimation. The proposed pipeline employs a piecewise aggregate approximation for dimensionality reduction, a singular spectrum analysis for noise suppression, and a central-moment-based change point detection to automatically identify the onset of degradation. Subsequently, a health index is constructed to characterize the degradation process and useful service life. An extreme gradient boosting classifier is adopted to distinguish between healthy and degraded operating states of the bearing using vibration response. Later, a hybrid convolution with gated recurrent (CGR) neural network is proposed to predict the RUL. The method is validated using multiple run-to-failure bearing datasets with different operating conditions. The corresponding results show that the classifier achieves average $F1$ score of approx. 0.993 with reliability indices >2.0, indicating a clear identification of health-state. The proposed CGR network exhibits attractive performance, with high regression accuracy in RUL prediction. The efficacy of the proposed network is also evaluated against state-of-the-art methods, including bidirectional long-short-term memory recurrent neural networks, gated

* Author to whom any correspondence should be addressed.

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

© 2026 The Author(s). Published by IOP Publishing Ltd on behalf of the Institute of Systems Engineering, China Academy of Engineering Physics

recurrent units, causal dilated convolution-based residual densenet with channel attention and temporal convolutional network with transformer. Overall, the results confirm that the proposed framework enables accurate, computationally efficient, and reliability-informed RUL predictions, making it suitable for real-time condition monitoring and predictive maintenance applications in rotating machinery.

Keywords: fault prognostics, reliability assessment of RUL predictions, vibration signal processing, deep learning

1. Introduction

Sudden bearing failure can cause unplanned downtime in rotating machinery and infrastructure. It requires proper maintenance and continuous health monitoring to ensure safety. Frequent maintenance of such critical systems often increases operational cost and hence, estimating the remaining service life becomes essential. Accurate prediction of remaining useful life (RUL) has become a potential challenge in prognostics and health management (PHM) to ensure better reliability [1–3]. Conventional techniques employ the estimation of statistical characteristics from vibration data based on the time, frequency, and time–frequency domains. These properties are widely referred as *features*, and are further correlated to determine the RUL of the bearings. Time-domain based features such as kurtosis and peak-to-peak values help detect bearing faults but often suffer from noise recorded in the vibration response [4]. However, the addition of frequency-domain based features usually aides the PHM process for better assessment. Lei *et al* [5] and Hu *et al* [6] employed ensemble empirical mode decomposition to estimate RUL of bearings. This has been further enhanced by incorporating the maximum kurtosis deconvolution approach [7]. Tobon-Mejia *et al* [8] developed a data-driven prognostic algorithm to evaluate RUL by combining hidden Markov model (HMM) and wavelet packet decomposition. The HMM has also been used with dynamic principal component analysis to evaluate bearing deterioration [9]. Similarly, an adaptive empirical wavelet transform technique has been considered to decompose vibration signals [10]. It is selected with the highest kurtosis value based on minimum entropy deconvolution with particle swarm optimization. This helps reduce errors [such as mean absolute error (MAE), root MAE (RMSE) etc] in the estimation of the bearing RUL. However, such approaches often lack robustness under in-field conditions due to noise sensitivity, scalability issues, and nonlinear degradation process.

The estimation of RUL has been coupled with deep learning (DL) models for the prediction and classification of the health-state for better performance [11]. Kankar *et al* [12] applied wavelet transformation with artificial neural network for the prediction of RUL. Soualhi *et al* [13] proposed using the support vector regression model with Hilbert–Huang transform for the RUL analysis of ball bearings. However, these methods proved to be less efficient and were later improved for the service life prediction using convolutional neural network (CNN) [14], its deep version (i.e. deep CNN) [15], and the hybrid version with recurrent neural network (RNN) [16]. Nemani *et al* [17] considered the long-short-term memory (LSTM) model which was further improved through a bi-directional approach (i.e. Bi-LSTM) [18] for early detection of bearing failures. This method assists in time series based training of the vibration signal, where the critical signal anomalies may remain undetected due to sensitivity of the threshold values.

Recently, multi-paradigm prognostics have gained increasing attention that decouples health-state identification and RUL regression. These frameworks typically involve a degradation-state detector or classifier to identify the onset of damage and subsequently, prediction of RUL [19–21]. These state-aware strategies have shown the capability to reduce false prognostic activation during healthy operation and enhance prediction stability under non-stationary operating conditions. Data-driven approaches to detect the change-point (CP) in vibration signals have been proposed for accurate identification of degradation onset. These approaches include autoencoder-based reconstruction error monitoring [22], central-moment variation analysis [23], kernel-based density change detection [24], and deep feature transition monitoring using CNN [25].

Prasad *et al* [26] used a piecewise aggregate approximation (PAA) to estimate the statistical moments. A sparse encoder was adopted for training purposes to identify the remaining serviceable life of rolling bearings. It was observed that the second central moment of the vibration data proves to be a good indicator of bearing health-state, i.e. normal, incipient, and severe. Gupta *et al* [27] further combined singular spectrum analysis (SSA) to reduce signal noise and used the gated recurrent unit (GRU) model for the prediction of bearing RUL. Li *et al* [28] suggested a causal dilated convolution based residual densely connected network (CARDenseNET) with channel attention. It helped improve the accuracy of the RUL prediction as a linear function using time series data. Cao *et al* [29] proposed a hybrid temporal convolutional network (TCN)-transformer architecture. The model integrates long-range temporal dependency learning from TCN with the multidimensional attention of the transformer to enable feature extraction from vibration data. However, continuous concatenation and processing of growing feature maps in the network increase computational effort that impacts its efficiency.

In the literature review, DL models such as CNN, LSTM, and hybrid deep networks have improved the accuracy of fault prognostics. However, they remain fundamentally

deterministic, and thus the RUL estimation often suffers uncertainty in the prediction due to the non-stationary model learning process, error, noise, and other unknown factors [30, 31]. This affects the suitability of the DL applications for safety-critical RUL predictions. Hence, uncertainty quantification in fault prognostics is required to assess reliability and improve decision-making of point-based RUL estimators in the PHM process [31, 32]. Sankararaman [30] quantified the uncertainty source in the ML model using Monte Carlo simulation (MCS) and Bayesian updating. Lin and Li [33] presented a Bayesian LSTM framework to predict RUL and quantify uncertainty using concrete dropout, heteroscedastic Gaussian modeling, and MCS. This framework has been applied to turbofan and battery datasets for RUL estimates under uncertainty. Although it resulted in better reliability, the error values (i.e. MAE and RMSE) were significantly high. Hu *et al* [34] introduced a multi-scale multi-frequency branch interactive spatio-temporal sequence prediction network (M^2BIST-SPNet) for RUL of railway signaling devices. The spatial and temporal features of the signal were trained through separate average pooling-based convolution networks integrated by cross-branches. Although the architecture focused on predictive accuracy, the reliability assessment and uncertainty quantification of the RUL estimates remain underexplored.

Xuan *et al* [35] proposed an uncertainty-aware CNN based RUL framework that combines mean–variance estimation with hyper-deep ensembles to jointly model aleatoric and epistemic uncertainties.Xu and Zio [36] further improved the DL model using split conformal prediction and Bayesian modeling to generate adaptive intervals for uncertainty estimation through variational inference. Uncertainty-aware prognostic frameworks employing deep ensembles [37], conformal prediction [38], and reliability-informed learning [39, 40] have highlighted the importance of calibrated prediction intervals alongside point estimates. These studies demonstrate that reliable interval prediction is essential for translating RUL forecasts into risk-aware maintenance decisions, particularly in safety-critical rotating machinery applications.

However, most of these uncertainty-aware methods have been primarily validated under controlled conditions, and their effectiveness for vibration-driven bearing systems operating in highly non-stationary, noise-contaminated environments remains underexplored. Moreover, these approaches typically address uncertainty at the prediction stage without explicitly incorporating degradation-onset detection and health-state-aware prognostic triggering, which may lead to premature or unreliable RUL estimates during healthy operation. Consequently, a unified vibration-driven framework is required that integrates uncertainty quantification, degradation-state identification, and reliability assessment for robust RUL prediction of roller bearings.

This study proposes to address the issue of uncertainty in RUL predictions through a health-state informed DL framework for bearing prognostics. In the first place, extreme gradient boosting (XGBoost) is adopted to classify operating conditions as healthy or degraded, enabling efficient screening with low computational cost. This is followed by a hybrid convolution with gated recurrent (CGR) neural network to estimate RUL. It combines CNN for feature extraction capability with GRU layers to strengthen sequential modeling. The vibration response is preprocessed using PAA for time step reduction and singular spectrum-based noise reduction. The framework also incorporates probability assessment to ensure that RUL predictions remain accurate and have better reliability under varying operating conditions. The resulting system forms a vibration-oriented unified PHM, which has potential for online applications. In brevity, the main contributions of this study can be summarized as follows:

- Development of a hybrid convolution gated RNN to improve the RUL prediction by leveraging spatial feature extraction and temporal sequence learning.
- Reliability assessment of DL models for bearing RUL predictions to ensure better and consistent performance at each time instance.
- Efficient state-aware prognostic activation framework for bearings under different operating conditions.

The remainder of this paper is structured as follows. Section 2 presents the problem formulation indicating vibration-signal pre-processing, feature extraction, and performance metrics. Section 3 describes the proposed bearing failure prognostic framework, details of the XGBoost-based health classification, and the proposed CGR-based RUL estimate. Section 4 discusses the estimation of error from the learning models and reliability assessment to check their performances. Section 5 reports the hyperparameter optimization and performance evaluation using multiple run-to-failure bearing datasets. Finally, section 6 provides the concluding remarks and outlines potential future research directions for the present work.

2. Problem statement

Fault prognosis often requires contact based measurement of responses such as strain, deformation, vibration, temperature, and load [41]. The present study focuses on vibration based responses measured using accelerometers as time series signals. These are further analyzed in the time, frequency and time-frequency domains, which provides data series in terms of time and frequency. The data series are assessed through statistical means to determine features, as explained below.

2.1. Feature extraction

In the time and frequency domains, statistical features such as mean, variance, skewness, and kurtosis are often considered in fault prognosis based on time and spectral data [4, 12]. Moreover, in time-series data additional features are also included, such as root mean square (RMS), peak-to-peak, shape factor, crest factor, impulse factor, and latitude

factor [9, 21]. These characteristic features capture complementary aspects of vibration signals, including energy distribution, impulsiveness, and spectral asymmetry for RUL evaluation. Often, these features are computed from the onset of degradation, called the condition CP. In this study, a central moment-based CP detection method is adopted. For a time-series signal y_j, the mth order central moment at scale s is computed from the segmented sequence as [26]

$$\gamma_m^s = \frac{s}{n}\sum_{k=1}^{n/s}(y_k^s - \bar{y}^s)^m \tag{1}$$

where k is the number of data points in each segment, $\bar{y}^s$ represents the mean of the segment, and n is the total number of data points. The factor s/n serves as a normalization constant that accounts for the number of segments for the chosen scale. Deviations in higher-order moments reflect distributional shifts in the vibration signals.

Let the mth central moment corresponding to the jth segment be denoted by $\gamma_{m,j}^s$, where each segment corresponds to the sampling time t_j. A slope-based threshold $\mathcal{T}$ is applied to detect the CP, marking the transition from steady operation to the degradation regime, which is expressed as

$$\mathcal{T} \leqslant \left(\frac{\gamma_{m,\,j-1}^s - \gamma_{m,1}^s}{t_{j-1} - t_1}\right)\cdot\left(\frac{t_j - t_{j-1}}{\gamma_{m,\,j}^s - \gamma_{m,\,j-1}^s}\right) \tag{2}$$

where t_j and t_{j-1} represent consecutive sampling instants corresponding to the jth and $(j-1)$th segments, respectively. The threshold $\mathcal{T} \in [1,6]$ is selected based on the sensitivity of the data [21]. The initiation of degradation is considered for the estimation of the h_j, which denotes the true normalized RUL at the time corresponding to the jth segment. Each segment is associated with a representative time instant t_j. The index is estimated using [27]

$$h_j = 1 - \frac{t_j - t_{\text{cp}}}{t_{\text{fail}} - t_{\text{cp}}}, \qquad t_j \geqslant t_{\text{cp}} \tag{3}$$

where t_{cp} denotes the detected CP and t_{fail} is the failure time. The index is defined between $[0,1]$, where unity means excellent condition and zero reflects complete bearing failure. The HI remains near unity during healthy operations and decreases monotonically after the CP occurrence. It provides a normalized representation of the bearing RUL based on the fraction of remaining life to the degradation duration. This index forms the basis for both classification and RUL prediction. Hence, its accuracy in the bearing fault prognosis is desirable. It can be noted that the above formulation requires knowledge of the failure time t_{fail} during supervised training. During prediction, the trained model can estimate the RUL directly from the observed vibration features without accessing future failure information. Thus, failure time is used only to construct training labels and does not introduce information leakage during inference. This can be helpful in real-time condition monitoring and predictive maintenance applications. The performance of the prediction models is often measured using different population based metrics based on actual and predicted outcomes. These metrics, as considered in this study, are discussed in detail below.

2.2. Performance metrics

The performance of classification training models is usually derived from the core components of a confusion matrix. These occurrences are—true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Assume that there are classes of data categorized as damaged and undamaged. Here, TP and TN represent the counts of damage and undamaged incidents that are correctly classified, respectively. If the models predicted the undamaged case as damaged, it is known as FP and vice versa, it is called FN. This helps in formulating widely accepted evaluation indices, such as precision, recall, accuracy and $F1$ score.

The precision $[\text{TP}/(\text{TP}+\text{FP})]$ reflects the ability of the ML model to avoid false alarms by ensuring that the predicted positives are indeed correct, whereas the recall $[\text{TP}/(\text{TP}+\text{FN})]$ focuses on identifying the ability of the model to minimize FN cases. Accuracy $[(\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\ \text{FP}+\text{FN})]$ measures the overall proportion of correct predictions, including both positive and negative outcomes. However, accuracy can be misleading, especially in datasets with imbalanced class distributions, where the model may achieve high accuracy by favoring the majority class. Further, the $F1$ score can be used to achieve a balanced measure that considers both precision and recall metrics. It is a harmonic mean of these metrics that is calculated as

$$\text{F1} = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{4}$$

A high $F1$ score indicates that the model performs well in terms of completeness and correctness, which is particularly critical in applications where FN and FP are equally undesirable. However, these metrics lack defining a reliability based index or score to assess the performance of the ML based classification.

In prognostics, statistical indices like MAE [27], mean bias error (MBE) [42], RMSE [28], coefficient of determination (R^2) [29] and adjusted coefficient of determination (R^2_{adj}) [29] are often adopted to quantitatively evaluate the performance of the predictive models. These performance metrics quantify the deviation between the true RUL values h and the predicted values $\hat{h}$, providing insights on the accuracy and reliability of the training models. The R^2_{adj} is estimated using

$$R^2_{\text{adj}} = 1 - \left(1 - R^2\right)\frac{N-1}{N-\mathbf{p}-1} \tag{5}$$

where $\mathbf{p}$ indicates the number of predictors and R is evaluated as $1 - \sum_{j=1}^{N}(h_j - \hat{h}_j)^2 / \sum_{j=1}^{N}(h_j - \bar{h})^2$. The aforementioned metrics help in error estimation and evaluate the performance of the model in an overall sense. However, bearing failure prognostics can be evaluated at different time instances. Thus, they lack in identifying the reliability of the prediction at any given time. This limits predictions to actionable RUL estimates that PHM practitioners can trust and take informed

decisions. In particular, it is preferred to have fault prognostics that – (i) robustly detect the onset of degradation, (ii) provide accurate point estimates of RUL, and (iii) quantify the uncertainty of those estimates so that maintenance decisions can incorporate risk. This study aims to address the above mentioned problem statement by reliability assessment of the prediction models and development of a reliable DL based prognostic framework tailored for roller bearing applications. This is discussed further in the following section.

3. CGR model for fault prognosis

The proposed DL-based prognostic offers classification and prediction of RUL in a multi-paradigm manner. Initially, the operating condition of the bearing is classified as either healthy or unhealthy. This marks the initiation of degradation and subsequently, the prediction of the residual service life. This modular approach ensures low computational cost during healthy operation, while enabling accurate prognostics in the degradation regime. Bearing vibration signals obtained from run-to-failure experiments are typically large, non-stationary, and potentially contaminated with noise. Training ML models directly on such raw signals often leads to instability and reduced predictive accuracy. Sequential pre-processing steps are employed to compress, smooth, and transform the signals into informative representations to address the challenges as explained further.

3.1. Vibration signal pre-processing

The dimensionality of the raw time series $\mathbf{X} = [x_1, \ldots, x_n]$ is reduced using PAA [43], which partitions the signal into l equal segments. It replaces each segment with the local mean value, which is expressed as

$$y_j^{(l)} = \frac{l}{n} \sum_{i=\frac{n}{l}(j-1)+1}^{\frac{n}{l}j} x_i, \quad j = 1, 2, \ldots, l. \tag{6}$$

The resulting compressed sequence $\mathbf{Y} = [y_1^{(l)}, \ldots, y_l^{(l)}]$ preserves global trends while reducing redundancy and suppressing high-frequency no . This reduction in dimensionality improves computational efficiency without missing the degradation information. The compressed sequence is further processed using SSA [44, 45]. The signal is embedded in a Hankel trajectory matrix and decomposed into eigentriples. A smooth reconstruction of the signal is obtained by retaining components associated with dominant eigenvalues. This effectively suppresses noise while retaining the degradation characteristics recorded in the vibration signal. This denoising step ensures that the subsequent extraction of features captures the true bearing health characteristics and is not influenced by noise. A flow of these steps to refine the feature extraction, CP detection, and HI formulation is illustrated in figure 1.

3.2. Classification of health

Once the signal is pre-processed, each sample is represented by a feature vector $\mathbf{z}_j$ which is inputted into XGBoost-based ML model for health-state classification. It is a scalable ensemble learning algorithm that constructs additive regression trees $\mathbf{F}_t$ to minimize a regularized objective function $\mathcal{L}$ as

$$\mathcal{L} = \sum_{j=1}^{N} L(y_j, \hat{y}_j) + \sum_{t=1}^{T} \Omega(\mathbf{F}_t). \tag{7}$$

In the above equation, $L(\cdot)$ is a differentiable loss function for classification, and $\Omega(\mathbf{F}_t)$ is the regularization term to control the complexity of tree. It is calculated using

$$\Omega(\mathbf{F}_t) = \Gamma\ell + \frac{1}{2}\lambda \sum_{i=1}^{\ell} w_i^2 \tag{8}$$

where ℓ is the number of leaves, w_i denotes the respective weight factor, Γ represents the complexity penalty, and λ is the L_2 regularization coefficient. The model is updated iteratively using gradient boosting as presented in figure 2, which follows

$$\hat{y}_j^{(t)} = \hat{y}_j^{(t-1)} + \eta \mathbf{F}_t(\mathbf{z}_j). \tag{9}$$

In equation (9), $\mathbf{F}_t(\mathbf{z}_j)$ denotes the output of the tth regression tree evaluated in the feature vector $\mathbf{z}_j$ with the learning rate η. The resulting boosted model is obtained as

$$\hat{y}_j = \sum_{t=1}^{T} \eta \mathbf{F}_t(\mathbf{z}_j). \tag{10}$$

For binary classification, the probability of degradation $\hat{p}_j$ can be obtained through a sigmoid function as [46]

$$\hat{p}_j = \sigma(\hat{y}_j) = \frac{1}{1 + \exp(-\hat{y}_j)}. \tag{11}$$

XGBoost is adopted in this study due to its robustness to noisy features, ability to handle heterogeneous inputs, and high classification accuracy with relatively low computational cost.

3.3. Prediction of RUL

Once the degrading health-state is detected, it indicates the presence of fault in the system. This activates a DL based fault prognostic model to estimate the normalized RUL. The proposed CGR architecture integrates the convolutional network for the extraction of spatial features and the iterative gated recurrent network for the temporal modeling, as shown in figure 3. Let $\mathbf{X}_j$ be the input feature sequence for time step j such that a convolutional layer applies a kernel matrix $\mathbf{W}$ to capture local degradation patterns as

$$\mathbf{S}_j = f(\mathbf{W} * \mathbf{X}_j + \mathbf{b}) \tag{12}$$

where $*$ denotes the convolution operator, $\mathbf{b}$ is the bias vector, and $f(\cdot)$ represents the rectified linear unit (ReLU) function. The pooling layers subsequently reduce the dimensionality while preserving salient information. The extracted feature maps are further processed by GRU to capture temporal dependencies. At each time step j, the reset gate is given by

$$\mathbf{r}_j = \sigma(\mathbf{W}_r[\mathbf{h}_{j-1}, \mathbf{S}_j] + \mathbf{b}_r), \tag{13}$$

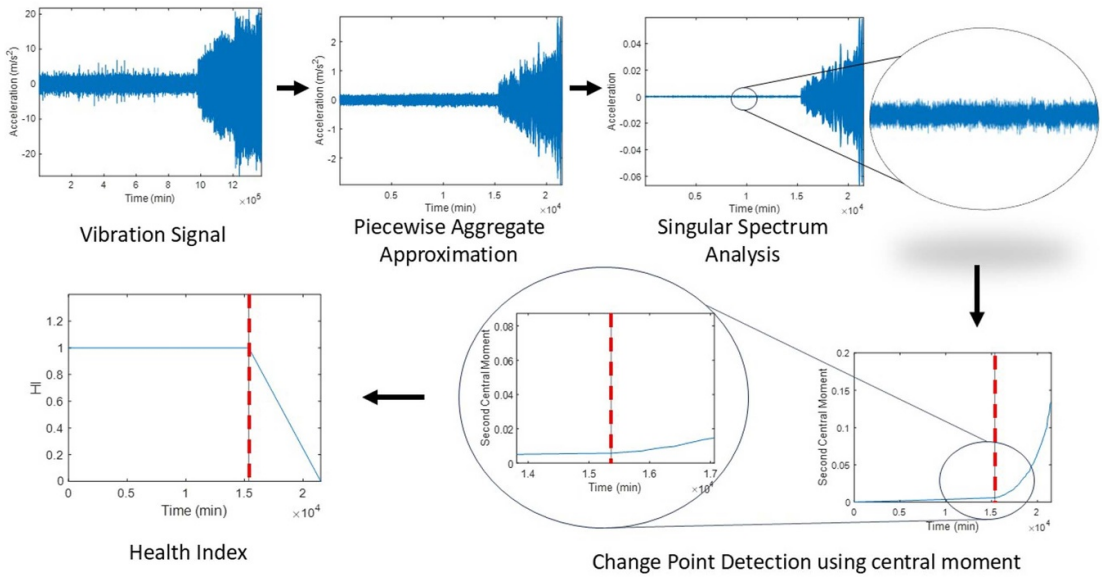

Figure 1. Pre-processing pipeline of vibration signals for feature extraction, change-point detection, and normalized RUL estimation.

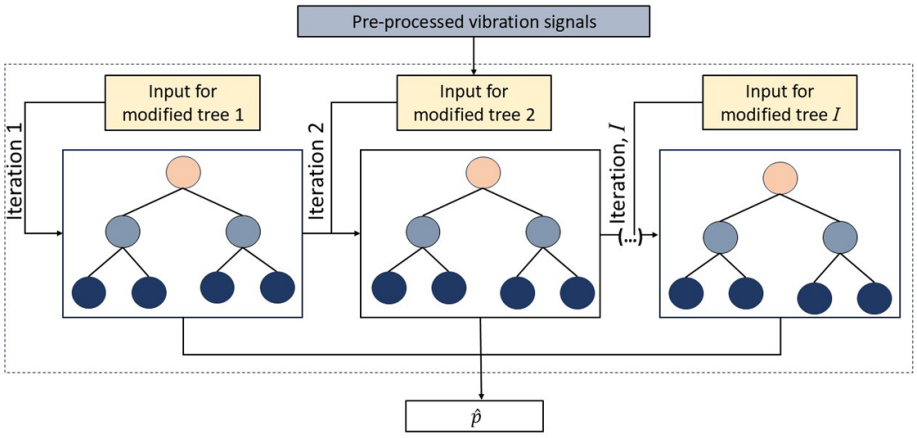

Figure 2. XGBoost-based bearing health-state classification model.

and the update gate follows

$$\mathbf{u}_j = \sigma\left(\mathbf{W}_u\left[\mathbf{h}_{j-1}, \mathbf{S}_j\right] + \mathbf{b}_u\right). \tag{14}$$

In equations (13) and (14), $\mathbf{h}_j$ represents the hidden state encoding in temporal context, which is expressed as

$$\mathbf{h}_j = (1 - \mathbf{u}_j) \odot \mathbf{h}_{j-1} + \mathbf{u}_j \odot \tilde{\mathbf{h}}_j \tag{15}$$

where $\odot$ denotes element-wise multiplication and candidate hidden state $\tilde{\mathbf{h}}_j = \tanh(\mathbf{W}_h[\mathbf{r}_j \odot \mathbf{h}_{j-1}, \mathbf{S}_j] + \mathbf{b}_h)$.

This step is performed in an iterative manner to improve the prediction accuracy. This enables for effective modeling of the evolution for bearing degradation. The final hidden state in the proposed CGR network is flattened and passed through two fully connected (FC) layers to generate a scalar normalized RUL estimate for the given time step.

The combined workflow of the proposed CGR network for classification and fault prognosis is summarized in figure 4. The process initiates with raw vibration signals from bearings that undergo pre-processing, feature extraction, and XGBoost based health-state classification. If any system degradation is detected using equation (2), the CGR architecture estimates the bearing RUL. Further, the performance of the learning models is evaluated using metrics as discussed in section 2.2. Additionally, reliability assessment is performed in this study to quantify the uncertainty and reliability of the prediction model for risk-aware prognostics applications.

4. Reliability assessment

In the present study, reliability analysis is performed for both classification and RUL predictions. Since the classification features binary dimensionality, the reliability assessment is formulated based on the probability of failure. The conventional estimation of the probability of failure p_f requires integrating unsatisfied cases [i.e. $g(\mathbf{u}) \leqslant 0$] as [47]

$$p_f = \int_{g(\mathbf{u})\leqslant 0} f_{\mathrm{U}}(\mathbf{u})\mathrm{d}\mathbf{u} \tag{16}$$

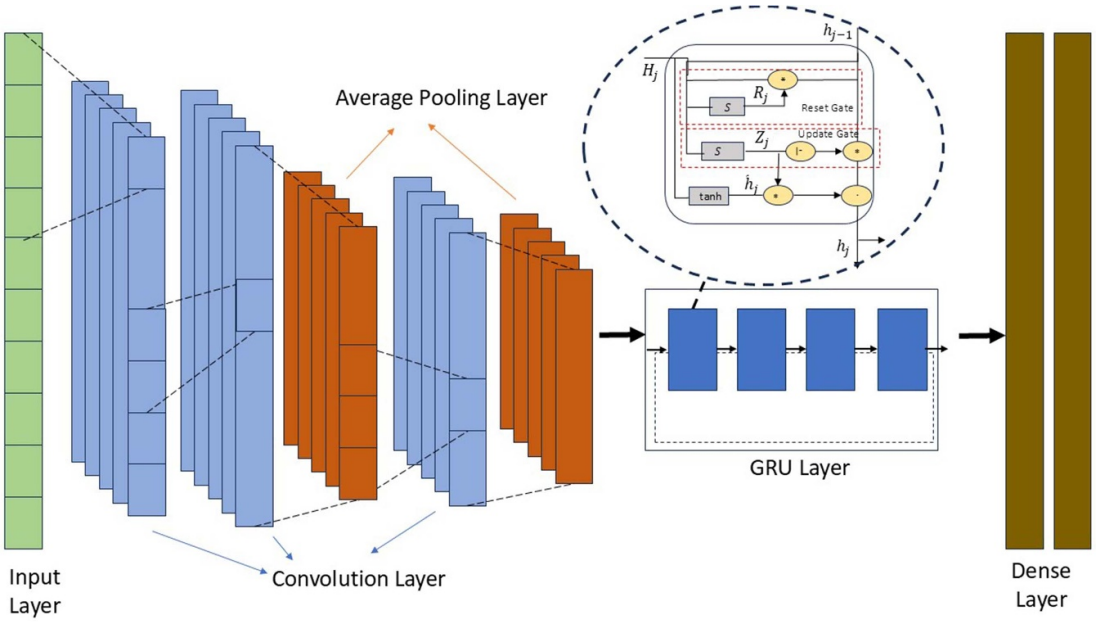

Figure 3. Architecture of the proposed CGR network for bearing RUL prediction.

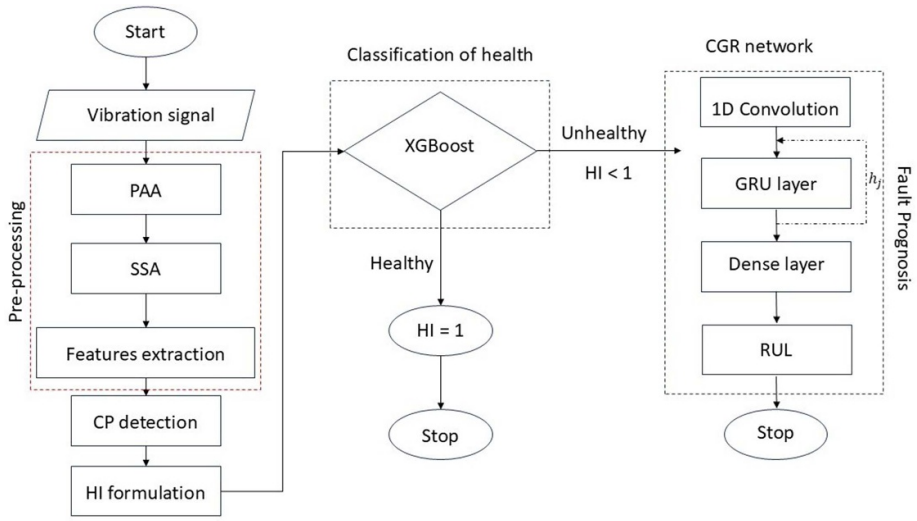

Figure 4. Schematic of the proposed CGR framework for classification and fault prognosis.

where $\mathbf{u}$ is the random variable in the function $g(\mathbf{u})$, and $f_{\mathbf{U}}(\mathbf{u})$ denote the probability density function. This formulation is discretized using population based estimation of the confusion matrix components. Here, the probability of misclassified predictions is assessed using

$$p_f = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{17}$$

The above expression indicates the failure of the classification model and thus, the failure corresponds to misclassifications in this context. Further, the reliability index β is obtained by mapping p_f to the standard normal space using [48]

$$\beta = -\Phi^{-1}(p_f) \tag{18}$$

where $\Phi^{-1}(\cdot)$ represents the inverse cumulative distribution function (CDF) of the standard normal distribution. It can be noted that index β is not related to the conventional structural reliability measure. A higher value of β indicates better classification reliability and therefore, less occurrence of misclassified cases.

Similarly, reliability is also assessed to quantify the accuracy in the RUL predictions from the DL models. The formulation estimates the proportion of predictions that do not exceed a predefined tolerance band, say $\pm\varepsilon$ [49]. This reflects the consistency of the model to produce reliable predictions rather than focusing solely on minimizing the magnitude of errors

in an average manner. The reliability R_l for the absolute prediction error (i.e. $|r_j| = |h_j - \hat{h}_j|$) is expressed as

$$R_l = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}\left(|r_j| \leqslant \varepsilon\right). \tag{19}$$

where N is the sample size of predictions. In the above equation, $\mathbb{I}(\cdot)$ denotes the indicator function, which gives unity when the condition is satisfied and otherwise yields zero. A value of R_l as 1 indicates that all predictions are within the acceptable tolerance range. This implies a reliable model, while R_l values close to zero suggest that the model frequently produces predictions outside of the acceptable error bounds. Thus, indicating poor reliability. In addition, the reliability of the prognostic intervals is also evaluated using the prediction interval coverage probability (PICP) [50] and the prediction interval normalized average width (PINAW) [51]. The PICP measures the proportion of true RUL values that fall within the predicted confidence interval as [50]

$$\text{PICP} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}\left(h_j \in [L_j, U_j]\right). \tag{20}$$

The prediction intervals for the RUL estimates are constructed using a non-parametric residual-based empirical quantile method. After obtaining the point prediction $\hat{h}_j$, the residual r_j is calculated at each time step. The pth quantile of the residual distribution q_p is determined as $\hat{F}_r^{-1}(p)$, where $\hat{F}_r(\cdot)$ denotes the empirical CDF of the residuals. The empirical CDF is used because the true distribution of residuals is unknown for the prediction models. Thus, the prediction confidence interval $[L_j, U_j]$ at significance level α (i.e. at $\alpha/2$ and $1-\alpha/2$ percentiles) is evaluated as $[\hat{h}_j + q_{\alpha/2}, \hat{h}_j + q_{1-\alpha/2}]$.

In PINAW based assessment, the average width of these intervals relative to the target variable range is evaluated using [51]

$$\text{PINAW} = \frac{1}{N(h_{\max} - h_{\min})}\sum_{j=1}^{N}(U_j - L_j) \tag{21}$$

where $h_{\max}$ and $h_{\min}$ represent the maximum and minimum values of h, respectively. A reliable model achieves high PICP with reasonably low PINAW, indicating accurate and limited uncertainty. The robustness of the prediction models can also be quantified using the Taylor skill score (TSS) given by [52]

$$\text{TSS} = \frac{4(1+c)}{\left(\frac{\sigma_{\hat{h}}}{\sigma_h} + \frac{\sigma_h}{\sigma_{\hat{h}}}\right)\left(\frac{\bar{\hat{h}}}{\bar{h}} + \frac{\bar{h}}{\bar{\hat{h}}}\right)}, \tag{22}$$

which integrates spatial correlation, bias, and variance into a unified metric. In the above equation, σ_h and $\sigma_{\hat{h}}$ are the standard deviations of the predicted values and the true observations, respectively, and c represents the spatial correlation coefficient. The TSS ranges between [0, 1], where the lower value indicates poor performance and vice versa. Overall, these metrics help to evaluate probabilistic accuracy measures, variance, uncertainty, error bias, and predictive reliability. It provides a comprehensive and rigorous evaluation of the RUL prediction models, which is demonstrated through experimental data in the next section.

5. Results

The proposed framework is validated using run-to-failure bearing vibration responses from the XJTU-SY and PRONOSTIA experimental dataset [53, 54]. The XJTU-SY dataset uses LDKUER204 type bearings with PCB 352C33 accelerometers to collect the vibration signals. The sampling frequency is 25.6 kHz, the sampling interval is 60 s, and the sampling time is 1.28 s. The dataset consists of 15 degradation-bearing signals under three operating conditions with different speeds and loads [53]. The details of the dataset nomenclature and operating conditions are further elaborated in table 1. In XJTU-SY dataset, the responses are recorded until failure, which gives life-cycle data with fault modes such as outer race, inner race, cage, and ball defects.

The PRONOSTIA dataset developed by the FEMTO-ST Institute consists of 17 run-to-failure bearing experiments conducted under three accelerated degradation operating conditions [54]. The experiments end when the acceleration amplitude exceeds 20g, providing complete life-cycle data from healthy operation to failure. The bearing operating conditions include speeds of 1800, 1650, and 1500 rpm with loads of 4000, 4200, and 5000 N, respectively. Based on these operating conditions, the nomenclature of the PRONOSTIA dataset is named B1_1, …, B1_7, B2_1, …, B2_7, B3_1, B3_2, and B3_3 in this study.

As suggested in section 3, the raw vibration signals are segmented into fixed-length windows using PAA. Each segment is filtered for noise reduction using SSA while preserving degradation trends. A total of 14 statistical features are evaluated including 10 features from time-domain and four features based on frequency-domain, as explained earlier, for each processed time segment. Figure 5 illustrates the normalized temporal evolution of a few vibration features such as peak, RMS, standard deviation, and spectral variance. Further, the CP is detected using the central moment formulation to mark the onset of bearing failure. A normalized degrading RUL is defined after the CP occurrence, which approaches zero at failure. This is used by the ML model to train the health-state classification. An ablation study is conducted based on XJTU-SY dataset to tune the model parameters and framework for better performance as suggested below.

5.1. Ablation study

In this section, multiple ablation studies are performed to assess the window size for signal segments, optimal hyperparameters for the classification model, roles of different modules in CGR framework, and DL model configurations for RUL. The vibration signal is divided into l segments based on window size as per equation (6). Different window sizes (ranging from 16 to 80) are considered to determine the optimal

Table 1. Operating conditions and lifetime of the XJTU-SY bearing vibration responses [53].

Dataset nomenclature	Operating condition		Bearing lifetime
	Rotating speed (rpm)	Load (kN)	
A1_1	2100	12	2 h 3 m
A1_2	2100	12	2 h 41 m
A1_3	2100	12	2 h 38 m
A1_4	2100	12	2 h 2 m
A1_5	2100	12	52 m
A2_1	2250	11	8 h 11 m
A2_2	2250	11	2 h 41 m
A2_3	2250	11	8 h 53 m
A2_4	2250	11	42 m
A2_5	2250	11	5 h 39 m
A3_1	2400	10	42 h 18 m
A3_2	2400	10	41 h 36 m
A3_3	2400	10	6 h 11 m
A3_4	2400	10	25 h 15 m
A3_5	2400	10	1 h 54 m

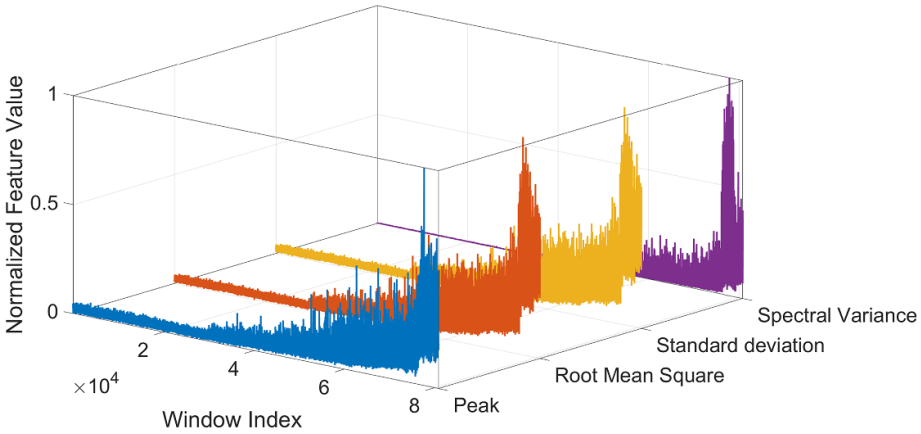

Figure 5. Illustration of the time- and frequency-domain vibration features of the XJTU-SY experimental bearing dataset.

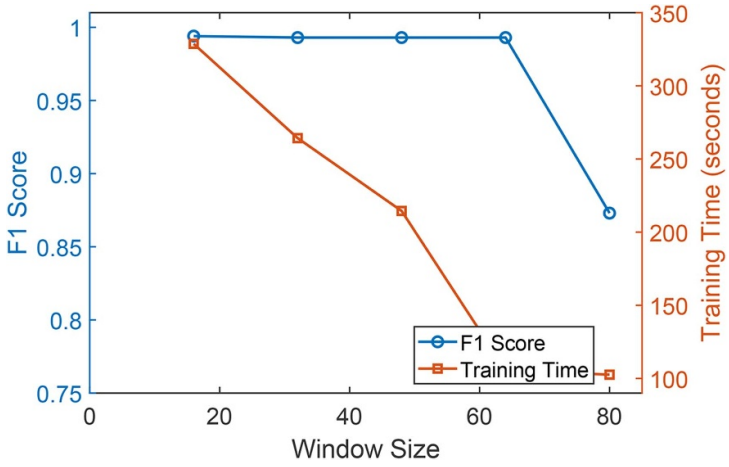

Figure 6. Comparison of classification model performance and training time with varying window sizes for signal segmentation.

size based on performance and efficiency. The present study adopts the XGBoost model for the classification of health-state to detect the initiation of damage. Figure 6 illustrates that the $F1$ score remains consistently high, about 0.993 across window sizes from 16 to 64, indicating stable classification performance. However, a significant drop in the $F1$ score is

Table 2. $F1$ scores obtained using grid search CV, random search CV, and Bayesian optimization methods for each bearing dataset.

Dataset	Grid search CV	Random Search CV	Bayesian optimization
A1_1	0.9890	0.9771	0.9773
A1_2	0.9758	0.9740	0.9742
A1_3	0.9969	0.9734	0.9755
A1_4	0.9996	0.9810	0.9888
A1_5	0.9993	0.9745	0.9755
A2_1	0.9970	0.9730	0.9765
A2_2	0.9859	0.9744	0.9781
A2_3	0.9847	0.9751	0.9786
A2_4	0.9995	0.9789	0.9879
A2_5	0.9916	0.9744	0.9801
A3_1	0.9914	0.9834	0.9892
A3_2	0.9914	0.9832	0.9868
A3_3	0.9998	0.9789	0.9875
A3_4	0.9958	0.9886	0.9905
A3_5	0.9976	0.9737	0.9864
Average	0.9930	0.9776	0.9823

observed at the window size of 80, suggesting a decrease in performance beyond the optimal range. The training time decreases monotonically as the window size increases. Hence, a window size of 64 provides a suitable trade-off between classification accuracy and computational efficiency, achieving near-maximum $F1$ score while significantly reducing training time. This value is adopted in the present study for further assessment of the model performance.

The optimal hyperparameters for the XGBoost model are further studied in this section using different search strategies—grid search cross-validation (CV), random search CV, and Bayesian optimization. The grid search CV method explores a structured and exhaustive combination of hyperparameter values. This results in an optimal configuration of the ML model with a maximum tree depth of 7, a learning rate of 0.1, an estimator size of 200, a minimum child weight of 3, and a gamma value of 0.5. The random search CV method optimizes the hyperparameter space in a stochastic manner. The method optimizes performance with a reduced maximum depth of 3, a relatively higher learning rate of 0.5, fewer estimators (i.e. 100), a minimum child weight of unity, and observes a decrease in the gamma value to 0.1. The last optimization method considered in this study, i.e. Bayesian optimization, adaptively updates the search strategy using probabilistic modeling. In this case, the optimal hyperparameter combinations for XGBoost are observed as maximum depth of 3, low learning rate of 1×10^{-2} with fewer estimators (i.e. 50), minimum child weight of 2, and gamma value of 0.5. These results highlight the differing search behaviors and their impact on the set of optimal hyperparameter configurations. The optimal hyperparameter configurations of different strategies are evaluated based on $F1$ scores as per equation (4). The results of the XJTU-SY bearing dataset are summarized in table 2. All strategies provided a near accurate $F1$ scores, however, the grid search CV method consistently achieves the highest scores. Based on this ablation study, the hyperparameters suggested by the grid search CV methods lead to an optimal model configuration for the bearing fault classification task. This configuration is adopted in the classification of health-state of bearings.

Further, the architectural sensitivity of the proposed CGR network is studied through a series of ablation configurations (i.e. case 1 to 10). This investigation systematically varies the number of convolutional layers, the number of GRUs, the learning rate, and the FC neurons as presented in table 3. Cases 1 and 2 examine the influence of shallower and deeper convolutional structures, respectively. Cases 3 and 4 evaluate the impact of reducing and increasing the number of GRUs. Cases 5 to 7 analyze the learning rate sensitivity, whereas cases 8 to 10 study the effect of decreasing and increasing the FC layer size. The results indicate that both under-parameterized (i.e. cases 1, 3 and 8) and over-parameterized (i.e. cases 2, 4 and 9) configurations degrade performance compared to the proposed setting, i.e. case 10. In particular, a large number of GRUs (in case 4) and higher learning rates (in case 6) lead to significant instability and reduced generalization capability of the DL model. The proposed CGR configuration (i.e. case 10 with 3 convolutional layers, 32 GRUs, learning rate 0.01, and 128 FC neurons) achieves the lowest MAE, lowest RMSE, and highest R^2 values. This demonstrates an optimal balance between representational capacity and training stability.

Additionally, the contribution of each module to the proposed CGR framework is investigated in this ablation study. These modules include pre-precessing (i.e. PAA and SSA) and ML models (i.e. XGBoost and CGR network) for classification and RUL prediction. Table 4 presents a systematic ablation study by progressively integrating PAA and SSA into the XGBoost and proposed CGR architecture to highlight their roles. The ML configuration without pre-processing shows comparatively weaker regression performance and stability. The inclusion of SSA improves noise robustness and predictive consistency, while PAA contributes to dimensionality reduction and better feature representation. Joint incorporation of both pre-processing modules demonstrates substantial improvement across classification and regression metrics. These findings confirm the efficacy and complementing role

Table 3. Ablation study of different CGR network configurations for bearing RUL predictions.

Cases	Conv. Layers	GRUs	Learning rate	FC Neurons	MAE	RMSE	R^2
Case 1	2	32	1×10^{-2}	128	0.096	0.126	0.809
Case 2	4	32	1×10^{-2}	128	0.096	0.125	0.814
Case 3	3	16	1×10^{-2}	128	0.094	0.123	0.818
Case 4	3	64	1×10^{-2}	128	0.181	0.231	0.359
Case 5	3	32	1×10^{-3}	128	0.091	0.122	0.822
Case 6	3	32	2×10^{-2}	128	0.250	0.289	−0.003
Case 7	3	64	2×10^{-2}	128	0.105	0.132	0.792
Case 8	3	32	1×10^{-2}	64	0.094	0.123	0.818
Case 9	3	32	1×10^{-2}	256	0.250	0.289	−0.001
Case 10	3	32	1×10^{-2}	128	0.006	0.008	0.999

Table 4. Ablation study of different module combinations under the proposed CGR framework.

PAA	SSA	XGBoost	CGR network	$F1$ Score	MAE	RMSE	R^2
—	—	✓	✓	0.989	0.241	0.335	0.450
—	✓	✓	✓	0.992	0.208	0.263	0.577
✓	—	✓	✓	0.991	0.199	0.247	0.725
✓	✓	✓	✓	0.993	0.006	0.008	0.999

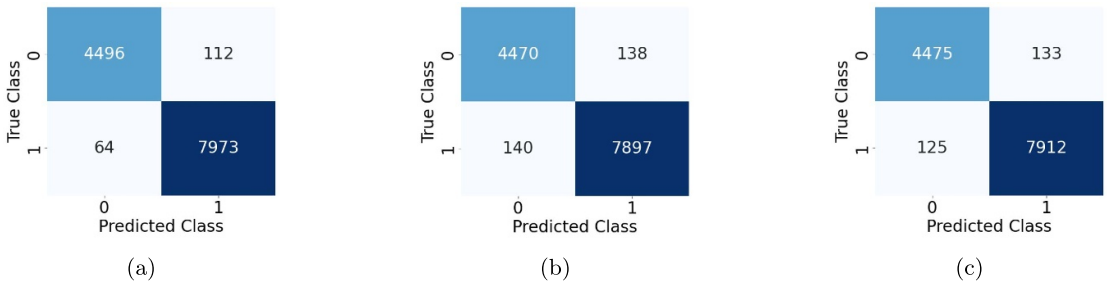

Figure 7. Health-state classification confusion matrices for A1_1 dataset using (a) grid search CV, (b) random search CV, and (c) Bayesian optimization.

of all modules considered in the proposed framework. Overall, the sequential and combined integration is essential to achieve optimal classification and predictive accuracy.

5.2. Health classification

The XGBoost classifier with grid search CV based optimized hyperparameters is trained to distinguish between healthy and degraded states of the XJTU-SY dataset. Figure 7(a) illustrates the confusion matrix for the A1_1 dataset to assess the classification performance of XGBoost. The model correctly identifies 7973 samples as belonging to the damaged class (i.e. TP) and 4496 samples as undamaged (i.e. TN). However, it suffers from a few misclassifications with 112 occurrences of FP and 64 time steps of FN. This number becomes larger in the case of hyperparameters tuned using the random search CV, and the Bayesian optimization methods as shown in figures 7(b) and (c), respectively. Moreover, the high concentration of predictions along the diagonal highlights the strong discriminative capability of the classifier. Figure 8 illustrates the t-distributed stochastic neighbor embedding (t-SNE) feature distribution for this dataset. The t-SNE plot reveals two well-separated clusters corresponding to the unhealthy (i.e. class 0) and healthy (i.e. class 1) conditions. This highlights the effective class-specific separability training of the XGBoost based model. The clear clustering behavior demonstrated by the proposed classifier strengthens the robustness of the health-state classification. Further, the performance of the XGBoost classifier is compared to other state-of-the-art ML models such as decision tree (DT), random forest (RF), naive Bayes, and support vector machine (SVM), as shown in table 5. Here, the optimal hyperparameters of the ML models are tuned based on the grid search CV. DT and RF based classifiers perform well with $F1$ scores of 0.9742 and 0.9889, respectively. However, it is marginally less than the performance of XGBoost model adopted in the present study. This strengthens the choice of the ML model for efficient classification of bearing health-state.

In this study, the reliability of classification based learning is computed using equation (18). The reliability index β is calculated for the classification of each XJTU-SY dataset as shown in table 6. It indicates a quantitative measure of the robustness of the classifier model under different bearing operating conditions. The highest probability of failure

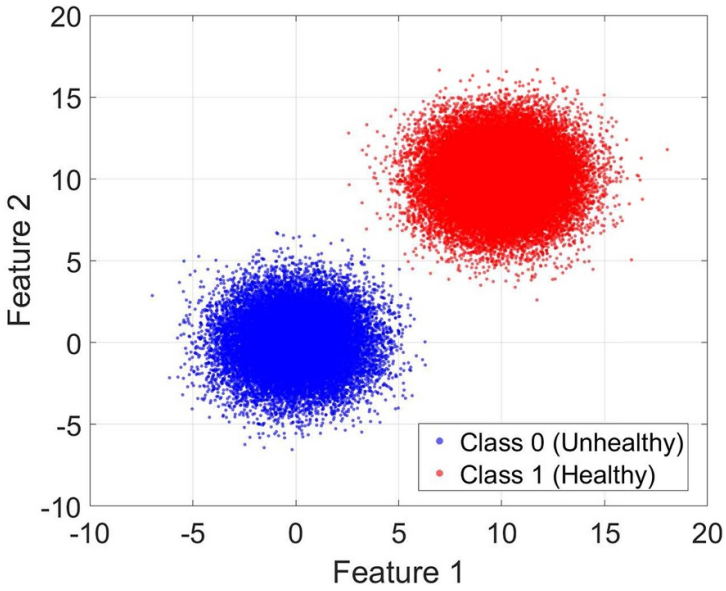

Figure 8. t-SNE visualization of the health-state classification of A1_1 dataset using XGBoost.

Table 5. $F1$ scores of different ML models hypertuned using grid search CV for classification of the bearing health using the XJTU-SY dataset.

Dataset	DT	RF	Naive bayes	SVM	XGBoost
A1_1	0.9634	0.9741	0.6392	0.8238	0.9890
A1_2	0.9594	0.9700	0.4410	0.8054	0.9758
A1_3	0.9734	0.9909	0.9660	0.9411	0.9969
A1_4	0.9967	0.9975	0.9966	0.9966	0.9996
A1_5	0.9694	0.9962	0.9671	0.7477	0.9993
A2_1	0.9942	0.9950	0.9268	0.9279	0.9970
A2_2	0.9660	0.9772	0.7048	0.7729	0.9859
A2_3	0.9564	0.9802	0.7707	0.8217	0.9847
A2_4	0.9870	0.9946	0.9484	0.7184	0.9995
A2_5	0.9679	0.9866	0.7714	0.6709	0.9916
A3_1	0.9547	0.9903	0.9847	0.6815	0.9914
A3_2	0.9550	0.9924	0.9487	0.7482	0.9914
A3_3	0.9959	0.9985	0.9268	0.9268	0.9998
A3_4	0.9882	0.9938	0.9516	0.8862	0.9958
A3_5	0.9851	0.9964	0.8512	0.9413	0.9976
Average	0.9742	0.9889	0.8530	0.8274	0.9930

p_f is observed to be 0.0192 and the lowest value is reported as 0.0003. Accordingly, the reliability index β is estimated between 2.0706 and 3.4316 for the XGBoost classifier using XJTU-SY dataset. The results correspond to a low misclassification probability by the ML model, and thus establishes the performance of the XGBoost classifier model in reliability sense too.

The PRONOSTIA bearing dataset is also considered for classification of health-state to further demonstrate the robustness and dataset independence of the proposed model. Table 7 presents the classification performance of the XGBoost model across different PRONOSTIA bearing degradation datasets. It is observed that the adopted classifier consistently results in a high precision, recall, accuracy, and $F1$ score with average values of 0.9905, 0.9966, 0.9891, and 0.9935, respectively. This indicates a strong discriminative capability, and well-balanced performance across varying operating conditions. The FN instances are observed to be low which ensure that degraded bearing conditions are not misclassified as healthy. This observation is critical in safety-oriented PHM applications. The highest probability of failure p_f is observed to be 0.0223 and the lowest value is reported as 0.0017. Also, the corresponding reliability index β of the classifier model remains consistently high with minimum and maximum values of 2.0084 and 2.9244, respectively. It reflects low misclassifications probability and stable degradation-state identification. The stable classification metrics across multiple accelerated degradation regimes further demonstrate the robustness and generalization capability of the XGBoost-based health-state identification within the proposed framework.

Table 6. Reliability index β computed for each bearing dataset based on misclassification.

Dataset	TP	TN	FP	FN	β
A1_1	7973	4496	112	64	2.2001
A1_2	6371	9799	235	81	2.0706
A1_3	15 139	947	94	1	2.5181
A1_4	12 439	44	9	1	3.1559
A1_5	3994	1325	6	0	3.0618
A2_1	46 496	3506	240	40	2.5364
A2_2	8601	7640	243	3	2.1727
A2_3	32 405	21 167	72	935	2.0858
A2_4	3072	1226	2	1	3.1947
A2_5	18 516	15 884	170	144	2.3656
A3_1	245 634	10 207	2500	1750	2.1370
A3_2	249 104	2186	1700	2600	2.1248
A3_3	35 231	2745	14	0	3.3528
A3_4	147 404	6492	940	300	2.4089
A3_5	880	10 789	3	1	3.4316

Table 7. Classification of health-state using XGBoost model on PRONOSTIA bearing dataset.

Dataset	TP	TN	FP	FN	Precision	Recall	Accuracy	$F1$ Score	p_f	β
B1_1	13 610	8314	450	50	0.9680	0.9963	0.9777	0.9820	0.0223	2.0084
B1_2	6515	304	120	31	0.9819	0.9953	0.9783	0.9885	0.0217	2.0198
B1_3	10 420	4680	193	107	0.9818	0.9898	0.9805	0.9858	0.0195	2.0646
B1_4	8690	2570	110	54	0.9875	0.9938	0.9856	0.9907	0.0144	2.1862
B1_5	19 050	620	27	7	0.9986	0.9996	0.9983	0.9991	0.0017	2.9244
B1_6	18 720	790	64	5	0.9966	0.9997	0.9965	0.9982	0.0035	2.6946
B1_7	17 440	590	31	11	0.9982	0.9994	0.9977	0.9988	0.0023	2.8305
B2_1	6900	285	17	86	0.9975	0.9877	0.9859	0.9926	0.0141	2.1936
B2_2	5985	355	24	12	0.9960	0.9980	0.9944	0.9970	0.0056	2.5335
B2_3	15 380	135	109	16	0.9930	0.9990	0.9920	0.9960	0.0080	2.4093
B2_4	5750	130	113	15	0.9807	0.9974	0.9787	0.9890	0.0213	2.0275
B2_5	17 930	510	41	7	0.9977	0.9996	0.9974	0.9987	0.0026	2.7948
B2_6	5410	183	15	5	0.9972	0.9991	0.9964	0.9982	0.0036	2.6909
B2_7	1768	42	24	6	0.9866	0.9966	0.9837	0.9916	0.0163	2.1369
B3_1	3875	205	30	10	0.9923	0.9974	0.9903	0.9949	0.0097	2.3374
B3_2	12 380	510	143	63	0.9886	0.9949	0.9843	0.9917	0.0157	2.1520
B3_3	2425	1035	7	5	0.9971	0.9979	0.9965	0.9975	0.0035	2.7010

5.3. Reliability assessment of RUL predictions

Once degradation is detected through the classifier, as explained earlier, the fault prognosis is activated to estimate the RUL of the bearings. The CGR network proposed in this study consists of three one-dimensional convolutional layers with a kernel size of three and a ReLU activation function. It has two average pooling layers for dimensionality reduction, which is followed by a GRU layer with 32 units to capture sequential dependencies. Finally, two FC layers of 128 neurons and 1 neuron are selected for regression output. The model is trained using Adam optimizer with a learning rate of 0.01, a batch size of 32, and 100 epochs. These hyperparameters are selected based on better prediction performance, and the proposed algorithm does not limit the choice of hyperparameters. The RUL prediction using the aforementioned hyperparameters and configuration is performed for all bearing cases used in this study.

Figure 9 represents the performance metrics of the proposed CGR network using XJTU-SY dataset. The average MAE, RMSE and R^2 values for the bearing dataset are observed to be 0.006, 0.008 and 0.999, respectively. Figure 10 illustrates the prediction of RUL of five bearings cases, *viz.* A1_3, A1_4, A1_5, A2_1, and A2_2 using the proposed CGR network. This estimation closely follows the actual bearing degradation trajectory after the occurrence of CP. The strong agreement in trend between the predicted and the true values of RUL supports the effectiveness of the proposed DL network. Moreover, in this study the prediction of RUL from other state-of-the-art methods such as CNN [14], LSTM [17], Bi-LSTM [18], GRU [27], CARDenseNet [28] and TCN-transformer [29] are compared as shown in figures 11–16, respectively. It is observed that LSTM, Bi-LSTM and CNN yield significant errors for the bearing dataset. This reflects a crucial PHM issue where the regression model of RUL prediction faces challenges despite a monotonic degradation due

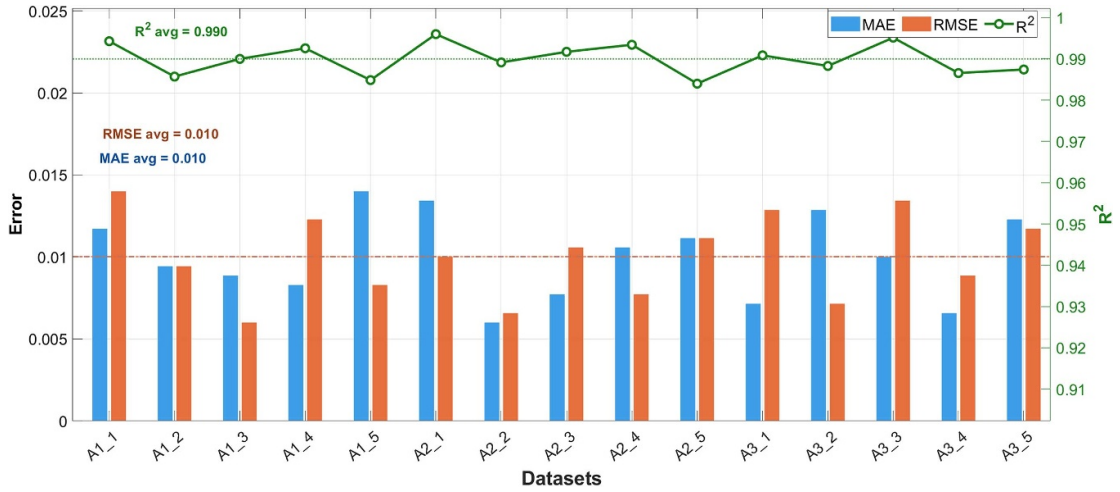

Figure 9. Performance comparison of the proposed CGR network in terms of MAE, RMSE, and R^2 for different bearing cases.

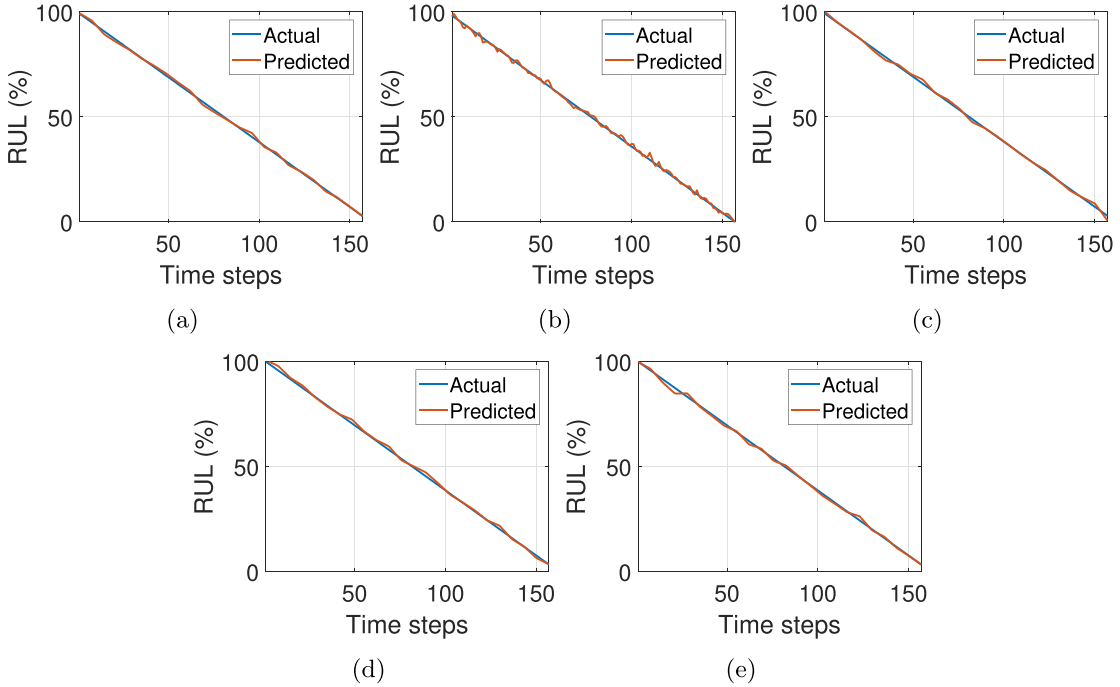

Figure 10. RUL predictions for (a) A1_3, (b) A1_4, (c) A1_5, (d) A2_1, and (e) A2_2 using the proposed hybrid CGR network.

to non-stationary characteristics of vibration signal. The prediction error is reduced by application of dense networks and transformer learning such as CARDenseNet [28] and TCN-transformer [29], however the proposed hybrid CGR performs better. This can be further summarized by estimating the statistical moments (i.e. mean, standard deviation, skewness, and kurtosis) of the error and absolute error, MAE, RMSE, MBE R^2, and R^2_{adj} as shown in table 8. The mean errors in the prediction of RUL from the DL networks considered in the study are often close to zero, indicating unbiased predictions. The magnitude of the mean error from LSTM, GRU, CARDenseNet, TCN-transformer and the proposed CGR is $\leqslant 0.01$. However, the standard deviations are high for the LSTM and GRU models. In case of CARDenseNet and TCN-transformer, the standard deviations of error are 0.041 and 0.038, respectively, which is higher than the proposed CGR network (i.e. 0.008). Similar observations can be noted in the case of absolute errors, here the proposed network performs better with a MAE of 0.006, and the standard deviation of absolute error is 0.005. Further, the CGR network achieves the lowest absolute error statistics (i.e. MAE is 0.006, RMSE is 0.008, and MBE is 0.001), and the prediction fits the actual RUL trend well with R^2 and R^2_{adj} of 0.999. This demonstrates a superior predictive accuracy and stability, which is better than CARDenseNet where the MAE, RMSE, and R^2 values are found to be 0.035, 0.045, and 0.975, respectively. The conventional recurrent

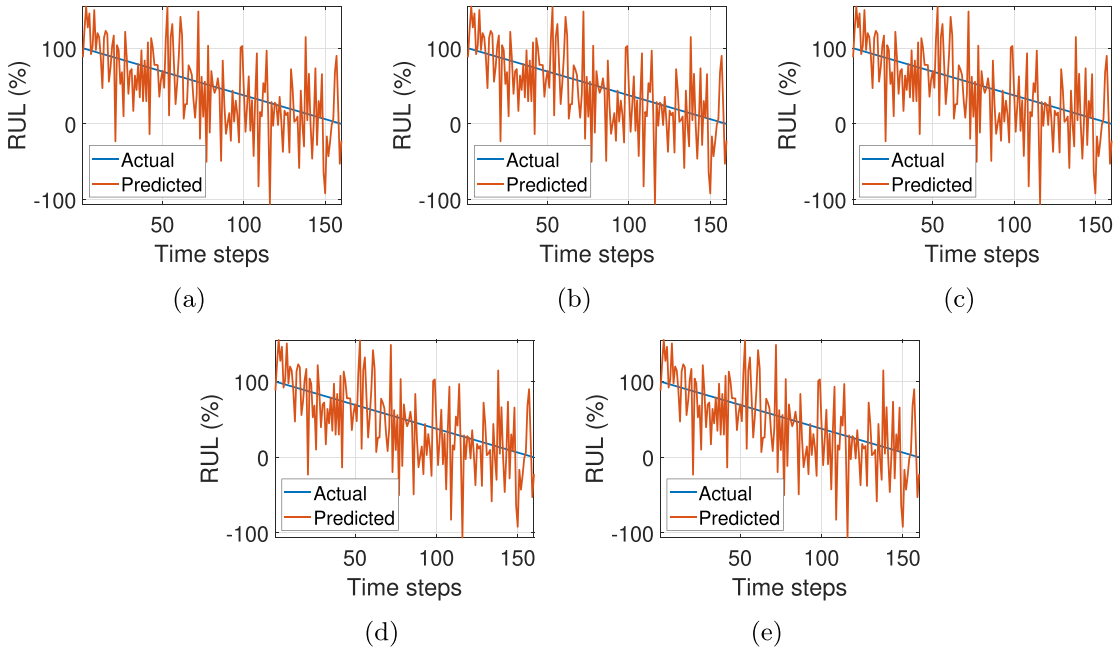

Figure 11. RUL predictions for (a) A1_3, (b) A1_4, (c) A1_5, (d) A2_1, and (e) A2_2 using CNN.

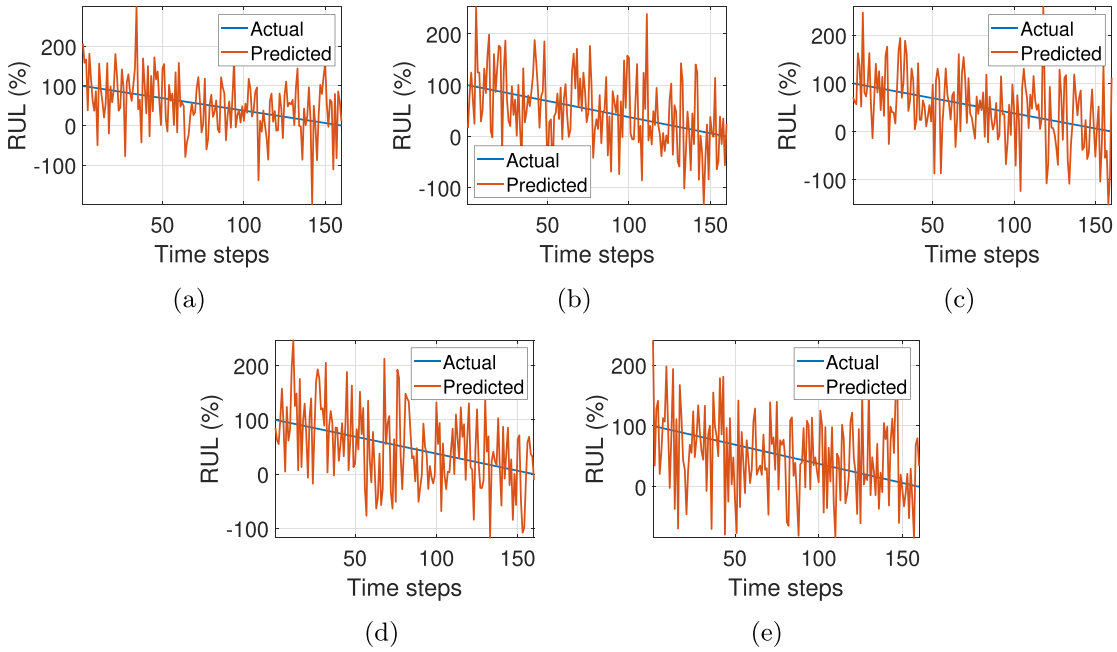

Figure 12. RUL predictions for (a) A1_3, (b) A1_4, (c) A1_5, (d) A2_1, and (e) A2_2 using LSTM model.

architectures (i.e. LSTM, Bi-LSTM and GRU) show a larger error dispersion, indicating a relatively poor fit on the evaluated dataset.

Further, the reliability assessment of the prediction models is performed based on different metrics as discussed in section 4. At first, the error in the prediction is calculated for each time step using true or actual RUL as discussed earlier. The probability mass function of the prediction errors is estimated for all bearing cases and DL models as adopted in this study. It provide the distribution of the prediction errors and illustrates model bias, if any. DL models with wider and asymmetric distributions indicate higher variability and directional bias in the predictions. Figure 17 presents the discrete probability distribution of error in the prediction of RUL for bearing A1_3 using different DL models. In consistency with the previous results, it is observed that variability and shift in the probability distributions are least in the case of the proposed CGR network compared to the other DL models used in this study. Additionally, uncertainty and reliability measures are also studied in this work to assess the performance of the DL model for bearing fault prognosis. These measures include PICP, PINAW, and TSS calculated using

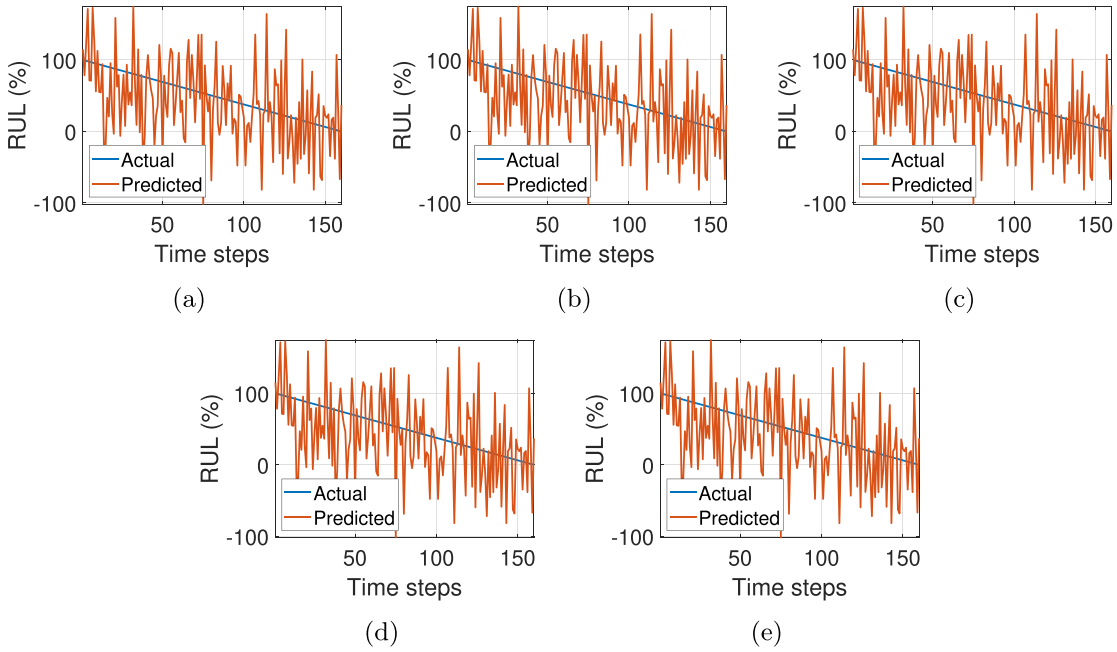

Figure 13. RUL predictions for (a) A1_3, (b) A1_4, (c) A1_5, (d) A2_1, and (e) A2_2 using Bi-LSTM model.

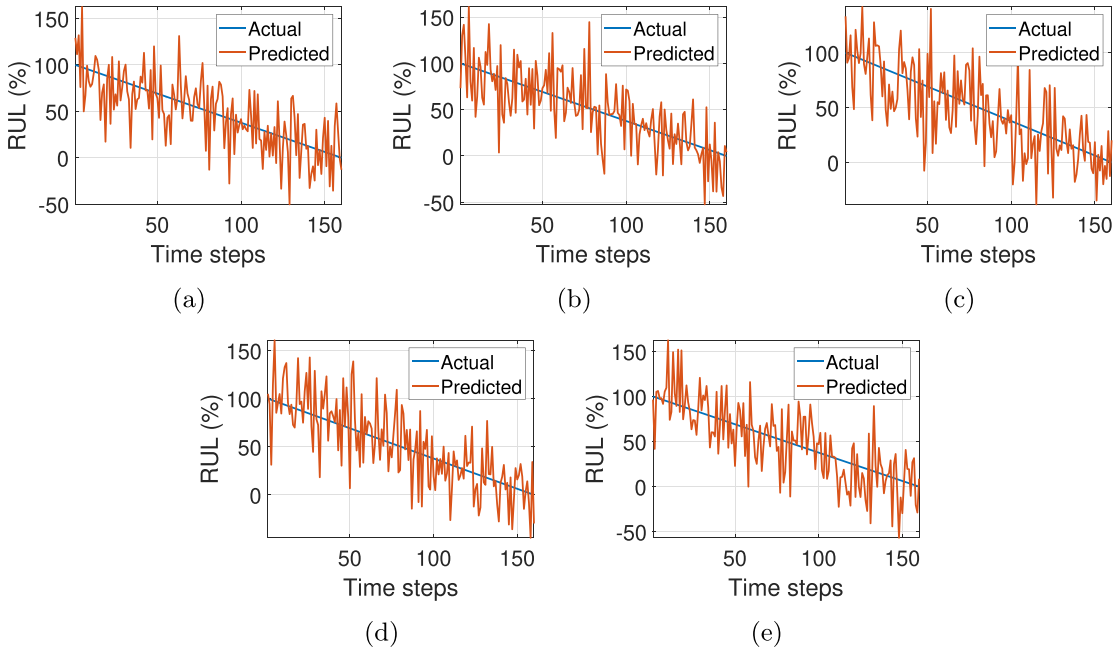

Figure 14. RUL predictions for (a) A1_3, (b) A1_4, (c) A1_5, (d) A2_1, and (e) A2_2 using GRU model.

equations (20)–(22), respectively. Consistent performance is reported for the proposed network based on PICP, PINAW, and TSS with values 0.950, 0.032, and 0.996, respectively, as presented in table 8. Other state-of-the-art models, such as CARDenseNet [28] and TCN-transformer [29], also perform fairly with comparable results in these reliability metrics, however the hybrid CGR network still yields better results. The other DL models used in this study, i.e. CNN, LSTM, Bi-LSTM, and GRU exhibit wider prediction intervals with larger PINAW values (i.e. >1) and low TSS values as reported in table 8.

Figure 18 illustrates the failure time steps of different DL models assuming a threshold ε of 0.02. In equation (19), these instances are noted as failure cases due to the exceedance of the limit state condition. This failure occurrence marks the poor performance of the RUL model prediction for the given time step, and it is not related to bearing failure. The threshold value can be suggested by the user as per the PHM operational requirement. For brevity, only five bearings cases (i.e. A1_3, A1_4, A1_5, A2_1, and A2_2) are presented in figure 18. It can be observed that CNN and recurrent architectures (i.e. Bi-LSTM, LSTM and GRU) have high

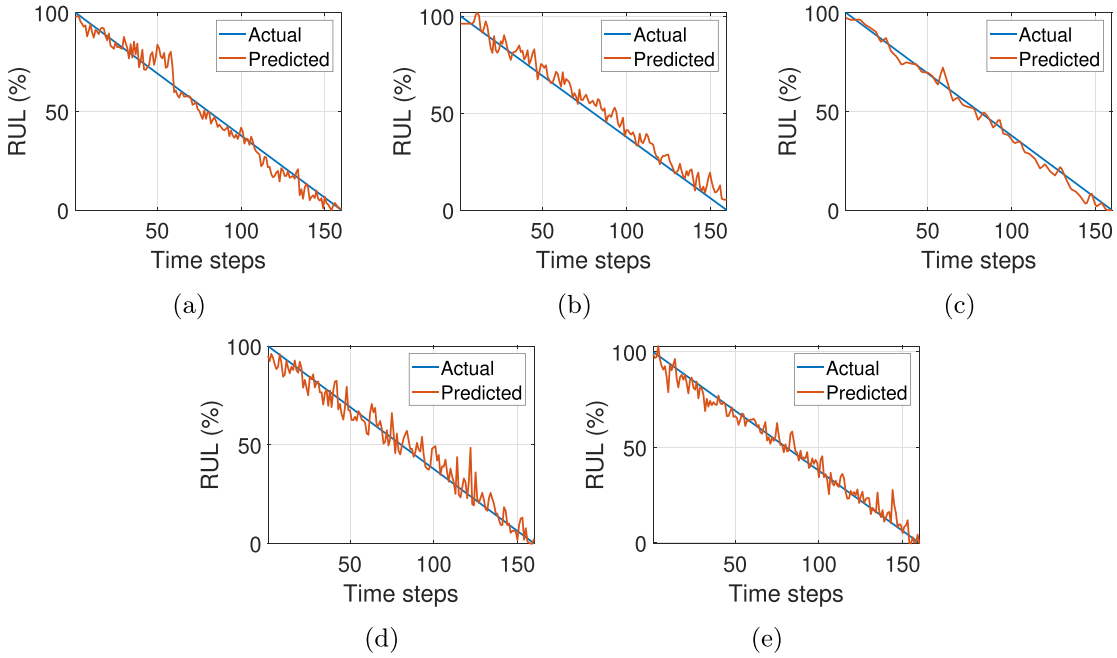

Figure 15. RUL predictions for (a) A1_3, (b) A1_4, (c) A1_5, (d) A2_1, and (e) A2_2 using CARDenseNet.

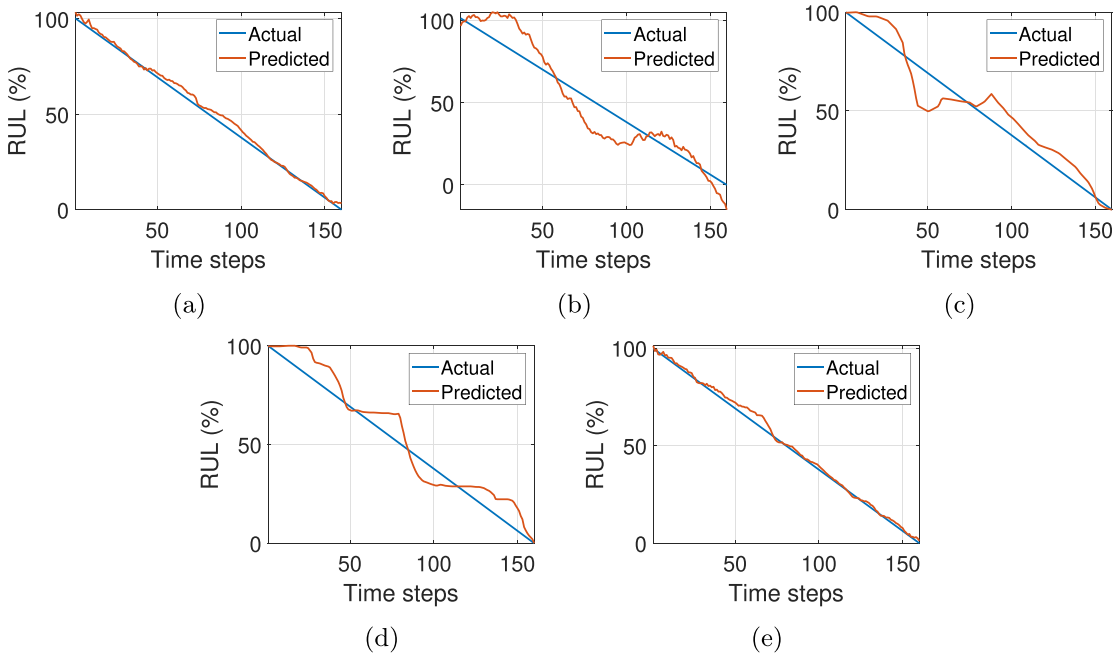

Figure 16. RUL predictions for (a) A1_3, (b) A1_4, (c) A1_5, (d) A2_1, and (e) A2_2 using the TCN-transformer.

variability, and failure cases are significantly away from the limit state. The CARDenseNet [28] and TCN-transformer [29] show limited variability from the limit state, while the proposed CGR network witnesses only a few failure cases. The observation can be further noted in the reliability assessment through varying threshold values as shown in figure 19. This reflects the reliability of the proposed method, which provides a low probability of failure in estimating the bearing prognostics. The results highlight the consistent behavior of the RUL predictions at different time instances through the proposed hybrid CGR network. It can be observed that the reliability of other DL networks such as CNN, LSTM, Bi-LSTM, and GRU is very low (i.e. 0.050, 0.040, 0.056, and 0.074, respectively, for $\varepsilon = 0.02$) for the bearing dataset considered (i.e. A1_3). Although, CARDenseNet [28] and TCN-transformer [29] improve the reliability to 0.380 and 0.487, respectively, but the proposed hybrid network outperforms it with a reliability value of 0.986 (for $\varepsilon = 0.02$). Overall, both point-prediction and interval-based metrics indicate that the proposed CGR network performs with

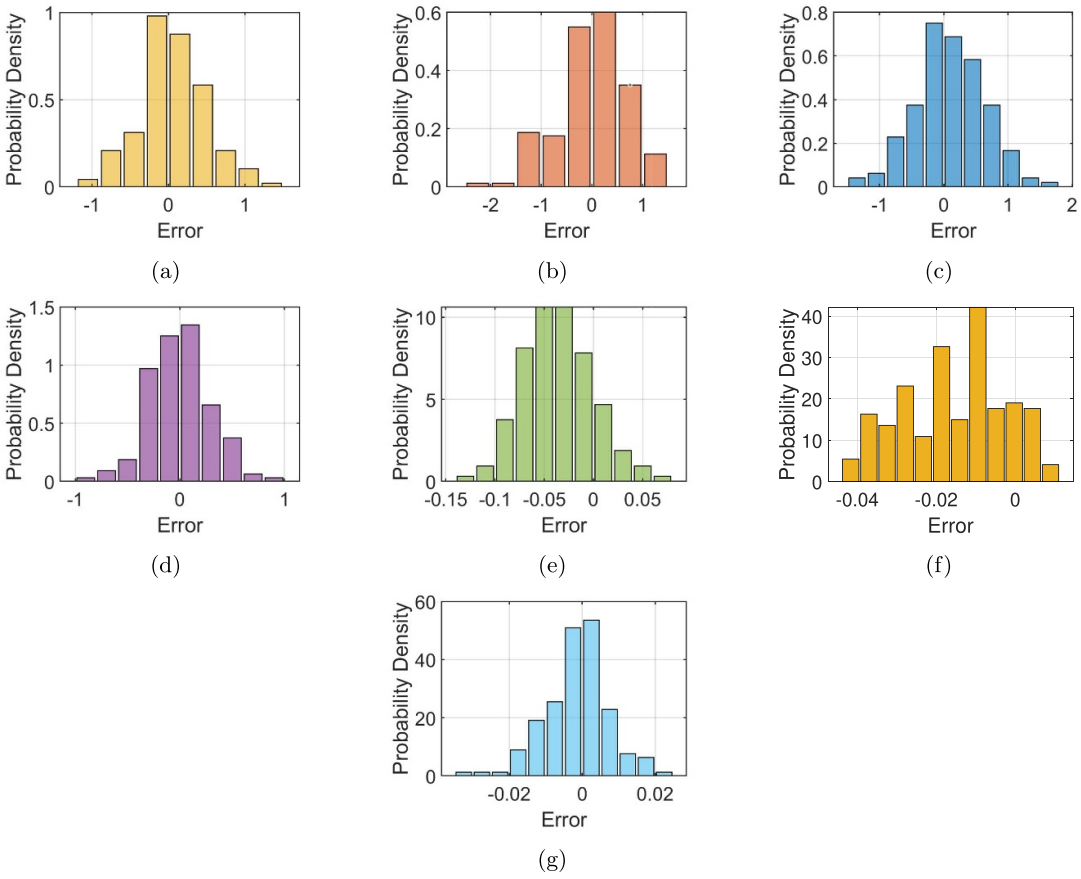

Figure 17. PDF of error in RUL prediction of A1_3 using (a) CNN, (b) LSTM, (c) Bi-LSTM, (d) GRU, (e) CARDenseNet, (f) TCN transformer and (g) the proposed CGR network.

Table 8. Summary of error statistics, regression metrics and uncertainty measures for the evaluated models.

Metric	Bi-LSTM	LSTM	CNN	GRU	CARDenseNet	TCN	Proposed CGR
Mean of error	0.104	−0.001	0.062	0.010	−0.004	−0.019	−0.001
Std. of error	0.533	0.661	0.437	0.281	0.041	0.038	0.008
Skewness of error	−0.051	−0.041	0.124	0.006	−0.566	0.297	0.032
Kurtosis of error	2.904	2.958	3.102	2.964	4.460	3.103	3.206
Mean of abs. error	0.429	0.530	0.340	0.220	0.035	0.037	0.006
Std. of abs. error	0.331	0.393	0.279	0.175	0.028	0.024	0.005
Skewness of abs. error	0.880	0.979	1.003	0.912	1.330	0.075	1.062
Kurtosis of abs. error	3.196	3.905	3.611	3.279	5.818	3.126	3.847
MAE	0.429	0.530	0.340	0.220	0.035	0.037	0.006
RMSE	0.541	0.659	0.440	0.281	0.045	0.055	0.008
R^2	−2.471	−4.152	−1.290	0.067	0.975	0.949	0.999
R^2_{adj}	−2.493	−4.184	−1.305	0.061	0.975	0.949	0.999
MBE	−0.104	0.001	−0.062	−0.010	0.004	0.001	0.001
PICP	0.876	0.905	0.877	0.907	0.950	0.945	0.950
PINAW	2.138	2.557	1.833	1.132	0.169	0.065	0.032
TSS	0.192	0.000	0.326	0.650	0.966	0.991	0.996

accurate and reliable RUL estimates among the evaluated models.

The above discussion establishes the superior performance of the proposed CGR network for a reliable assessment of bearing prognostics. Figure 20 illustrates the reliability assessment of the bearings considered in the XJTU-SY dataset. Different thresholds (i.e. ε ranging from 0.01 to 0.04) are considered for the construction of the limit state and reliability assessment. This marks a critical outcome of the study that provides the probabilistic significance of the bearing service life prediction by DL models and hence, helps in decision making of PHM.

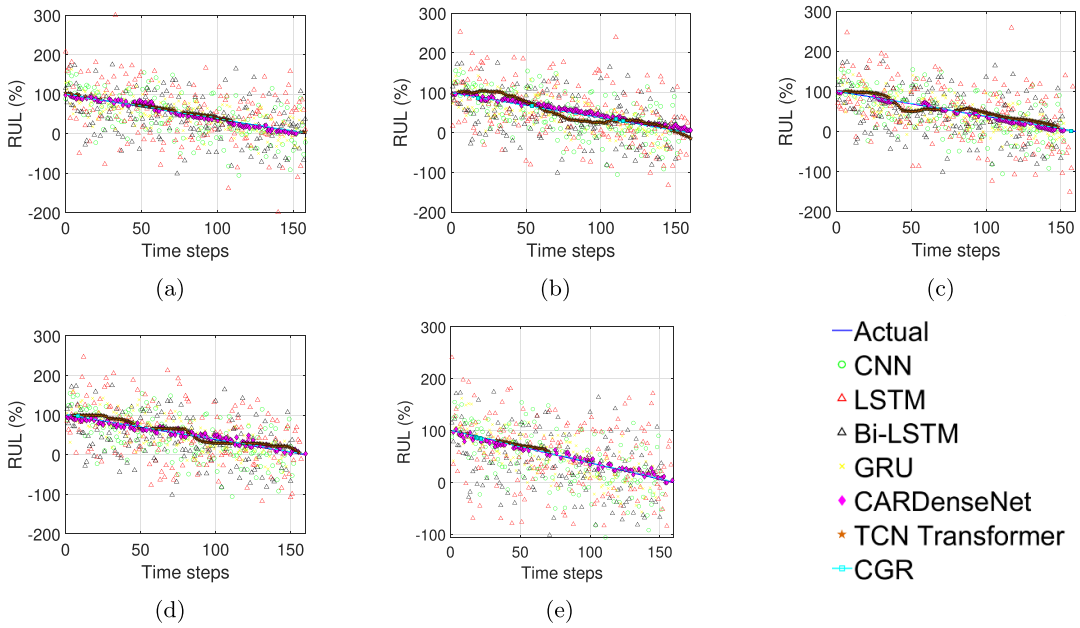

Figure 18. Illustration of the RUL prediction errors beyond the threshold value using different networks for (a) A1_3, (b) A1_4, (c) A1_5, (d) A2_1 and (e) A2_2.

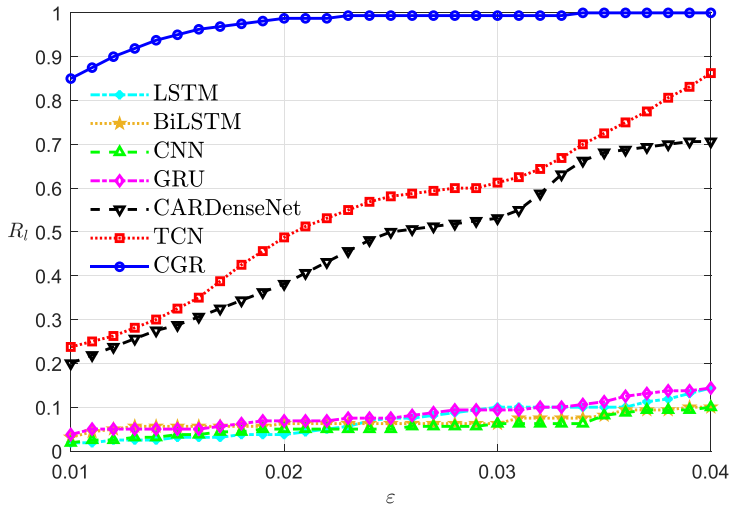

Figure 19. Reliability assessment of the DL models for A1_3 bearing RUL predictions at different error thresholds ε.

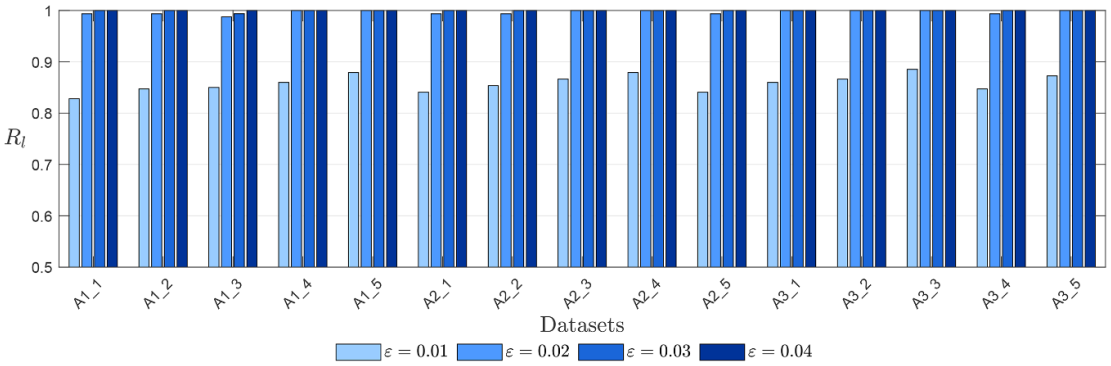

Figure 20. Reliability assessment of RUL predictions of XJTU-SY dataset bearings using the proposed hybrid CGR network.

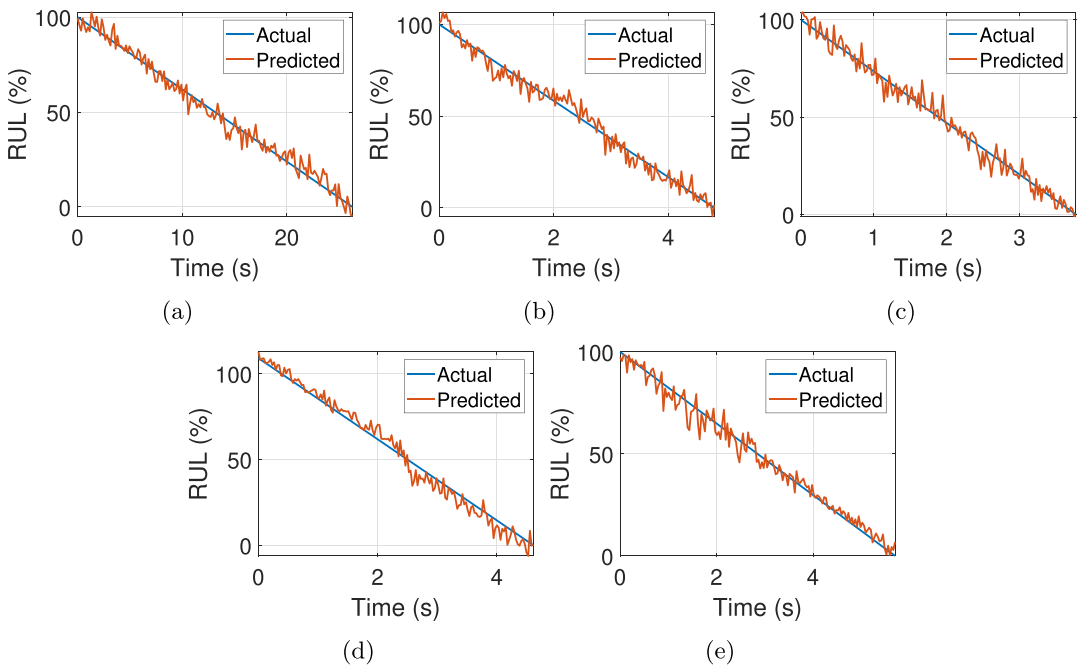

Figure 21. RUL predictions for (a) B1_1, (b) B1_2, (c) B2_1, (d) B2_2, and (e) B3_2 of PRONOSTIA dataset using the proposed hybrid CGR network.

Table 9. Performance metrics of the RUL predictions based on the proposed CGR network using PRONOSTIA bearing dataset.

Bearings	Mean err.	Std err.	Skew err.	Kurt err.	MAE	RMSE	R^2	Adj R^2	MBE	PICP	PINAW
B1_1	−0.002 674	0.041 745	−0.0916	3.0058	0.033	0.042	0.979	0.979	0.0027	0.95	0.1680
B1_2	−0.000 659	0.044 016	0.1430	2.9307	0.035	0.045	0.977	0.977	0.0007	0.95	0.1706
B1_3	−0.005 674	0.045 360	0.2421	3.2860	0.036	0.045	0.964	0.964	0.0035	0.95	0.1630
B1_4	0.000 842	0.042 731	0.2284	3.3016	0.034	0.044	0.977	0.977	0.0020	0.95	0.1720
B1_5	−0.000 517	0.043 988	0.2567	3.2699	0.034	0.046	0.965	0.965	0.0020	0.95	0.1740
B1_6	0.000 731	0.042 665	0.2420	3.2947	0.036	0.047	0.975	0.975	0.0035	0.95	0.1680
B1_7	0.000 468	0.043 421	0.2335	3.2781	0.035	0.045	0.976	0.976	0.0020	0.95	0.1790
B2_1	−0.000 911	0.043 142	−0.0725	3.6698	0.033	0.040	0.978	0.977	0.0009	0.95	0.1795
B2_2	−0.003 522	0.045 360	0.3972	2.7607	0.036	0.045	0.979	0.979	0.0035	0.95	0.1649
B2_3	0.000 902	0.042 889	0.2476	3.2924	0.033	0.042	0.980	0.980	0.0010	0.95	0.1660
B2_4	0.000 389	0.043 601	0.2388	3.2811	0.034	0.042	0.975	0.975	0.0020	0.95	0.1760
B2_5	−0.000 714	0.042 998	0.2449	3.2969	0.035	0.045	0.980	0.980	0.0035	0.95	0.1800
B2_6	−0.000 553	0.043 311	0.2366	3.2844	0.036	0.045	0.970	0.970	0.0010	0.95	0.1660
B2_7	0.000 668	0.043 044	0.2417	3.2888	0.032	0.040	0.970	0.970	0.0010	0.95	0.1690
B3_1	−0.000 581	0.043 277	0.2391	3.2840	0.031	0.040	0.979	0.979	0.0020	0.95	0.1770
B3_2	0.002 535	0.041 496	0.8344	4.0630	0.032	0.040	0.980	0.980	0.0010	0.95	0.1817
B3_3	0.000 623	0.043 109	0.2432	3.2894	0.033	0.043	0.970	0.970	0.0010	0.95	0.1760
Average	−0.000 509	0.043 303	0.2415	3.2869	0.034	0.043	0.975	0.975	0.0020	0.95	0.1724

A similar exercise is performed for the 17 bearings in the PRONOSTIA dataset. The RUL predictions are performed once degradation in the bearing vibration signal is detected through the classifier. The architecture and hyperparameters of the proposed hybrid CGR network remain unchanged as discussed earlier. Figure 21 illustrates the RUL predictions of the PRONOSTIA dataset using the proposed network. In concise, prediction of five bearing cases (i.e. B1_1, B1_2, B2_1, B2_2, and B3_2) are shown here to establish the merit. The estimation follows a similar performance with the actual bearing service life trajectory. The statistical outcomes from the RUL predictions of all the bearing under this dataset are summarized in table 9. The average values of the metrics show low scores of error (i.e. MAE = 0.034, RMSE = 0.043, and MBE = 0.002) and high values of R^2, and R^2_{adj} (i.e. 0.975). The reliability metrics such as PICP and PINAW also yield decent performance with scores of 0.95 and 0.17, respectively. These results demonstrate a strong agreement with outcomes observed earlier compared to the state-of-the-art networks considered in this study. Further, the results of the proposed CGR network are compared with the TCN-transformer based on the previous performance. Reliability assessment is performed for B1_1

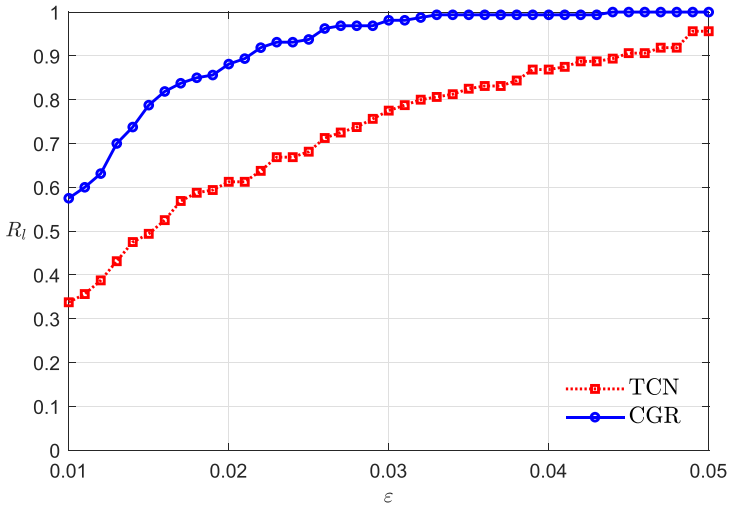

Figure 22. Reliability assessment of the DL models for B1_1 bearing RUL predictions at different error thresholds ε.

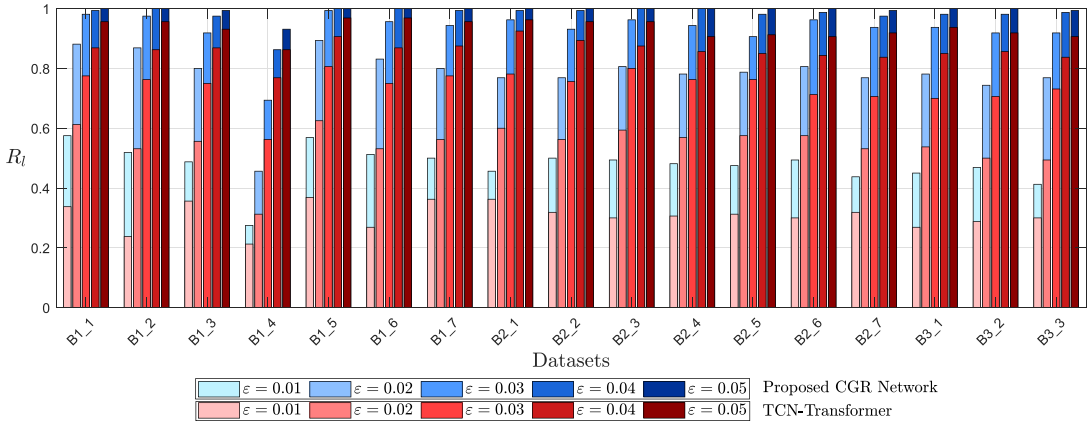

Figure 23. Reliability assessment of RUL predictions of PRONOSTIA dataset bearings using the proposed hybrid CGR network and TCN-transformer model.

bearing case with different thresholds as presented in figure 22. The results clearly delineate the superiority of the proposed DL model in providing reliable RUL predictions. This observation can be further reinforced for all other bearings as shown in figure 23. It highlights the significant gain in the reliability of RUL predictions for all bearings in this dataset. The assessment of the two datasets with different operating conditions tests the robustness and efficacy of the proposed hybrid CGR framework.

6. Conclusion

This paper presents a reliability assessment of prognostic frameworks for roller bearings. Additionally, a DL framework is proposed that combines robust signal pre-processing, XGBoost-based health-state classifier, and a hybrid convolution with gated recurrent network for RUL prediction with uncertainty quantification. It follows a PAA compression before singular spectrum-based denoising, and central moment-based change point detection. Furthermore, the signal features in the statistical sense are evaluated based on time- and frequency-domains, which act as inputs for training the classifier and DL network. XGBoost classifier is adopted for low-cost monitoring and timely detection of health degradation. It yields consistent high classification performance with an average $F1$ score of approx. 0.993, and reliability index of $>$2.0. This indicates a low misclassification probability and better reliability in identifying the health-state of the bearings. Once unhealthy, a hybrid architecture using convolution with iterative gated RNN (i.e. CGR) is proposed to produce point and interval estimates of bearing service life. The proposal is tested on multiple run-to-failure bearing datasets with different operating conditions to demonstrate the effectiveness of the approach and reliability assessment with state-of-the-art prediction models such as CNN, LSTM, Bi-LSTM, GRU, CARDenseNet, and TCN-transformer. In RUL predictions, the proposed CGR network performs better under point-prediction accuracy with MAE of 0.006, RMSE of 0.008, and R^2 of 0.999 for XJYU-SY bearing dataset. This efficacy is reflected in the probabilistic assessment with PICP of 0.95, PINAW of 0.032, and TSS of 0.996. In reliability assessment, the proposed network obtains significantly better reliability

in RUL predictions which outperforms the other DL models considered in the study. The study highlights the usage of probabilistic-based metrics to enhance reliability of DL models for accurate predictions at each time step.

Beyond performance improvements, the framework emphasizes computational prudence by invoking the computationally intensive RUL estimator only when degradation is detected, making it suitable for online or embedded monitoring systems. The future work will focus on integrating online learning, domain-adaptation techniques to handle changing operating conditions, and coupling prognostics with cost-based decision models to translate probabilistic RUL forecasts into optimized PHM policies.

ORCID iDs

Santosh Bisoyi 0009-0004-1609-6518
Amit Kumar Rathi 0000-0002-6193-8903
Swarup Mahato 0000-0002-6405-8768

References

[1] Jardine A K S, Lin D and Banjevic D 2006 A review on machinery diagnostics and prognostics implementing condition-based maintenance *Mech. Syst. Signal Process.* **20** 1483–510
[2] Huang R, Xi L, Li X, Liu C R, Qiu H and Lee J 2007 Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods *Mech. Syst. Signal Process.* **21** 193–207
[3] Zio E 2022 Prognostics and Health Management (PHM): where are we and where do we (need to) go in theory and practice *Reliab. Eng. Syst. Saf.* **218** 108119
[4] Saidi L, Ben Ali J, Bechhoefer E and Benbouzid M 2017 Wind turbine high-speed shaft bearings health prognosis through a spectral Kurtosis-derived indices and SVR *Appl. Acoust.* **120** 1–8
[5] Lei Y, He Z and Zi Y 2011 EEMD method and WNN for fault diagnosis of locomotive roller bearings *Expert Syst. Appl.* **38** 7334–41
[6] Hu L, Wang L, Chen Y, Hu N and Jiang Y 2022 Bearing fault diagnosis using piecewise aggregate approximation and complete ensemble empirical mode decomposition with adaptive noise *Sensors* **22** 6599
[7] Liu H, Zhou J, Zheng Y, Jiang W and Zhang Y 2018 Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders *ISA Trans.* **77** 167–78
[8] Tobon-Mejia D A, Medjaher K, Zerhouni N and Tripot G 2012 A data-driven failure prognostics method based on mixture of gaussians hidden markov models *IEEE Trans. Reliab.* **61** 491–503
[9] Yu J 2012 Health condition monitoring of machines based on hidden Markov model and contribution analysis *IEEE Trans. Instrum. Meas.* **61** 2200–11
[10] McDonald G L and Zhao Q 2017 Multipoint optimal minimum entropy deconvolution and convolution fix: application to vibration fault detection *Mech. Syst. Signal Process.* **82** 461–77
[11] Liu R, Yang B, Zio E and Chen X 2018 Artificial intelligence for fault diagnosis of rotating machinery: a review *Mech. Syst. Signal Process.* **108** 33–47
[12] Kankar P K, Sharma S C and Harsha S P 2011 Rolling element bearing fault diagnosis using wavelet transform *Neurocomputing* **74** 1638–45
[13] Soualhi A, Medjaher K and Zerhouni N 2015 Bearing health monitoring based on Hilbert–Huang transform, support vector machine and regression *IEEE Trans. Instrum. Meas.* **64** 52–62
[14] Eren L, Ince T and Kiranyaz S 2019 A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier *J. Signal Process. Syst.* **91** 179–89
[15] Star M and McKee K 2021 Remaining useful life estimation using neural ordinary differential equations *Int. J. Progn. Health Manage.* **12** 1
[16] Li J, Wang Z, Liu X and Feng Z 2023 Remaining useful life prediction of rolling bearings using GRU-DeepAR with adaptive failure threshold *Sensors* **23** 1144
[17] Nemani V P, Lu H, Thelen A, Hu C and Zimmerman A T 2022 Ensembles of probabilistic LSTM predictors and correctors for bearing prognostics using industrial standards *Neurocomputing* **491** 575–96
[18] Rathore M S and Harsha S P 2022 Prognostics analysis of rolling bearing based on bi-directional LSTM and attention mechanism *J. Fail. Anal. Prev.* **22** 704–23
[19] Sun B, Zhang J, Hu W, Wei Y, Wu X and Zhang L 2026 Prediction of remaining useful life for rolling bearings based on DCNN-BiLSTM-CBAM with multi-scale feature extraction *Eng. Res. Express* **8** 025513
[20] Guo J, Zhou J, Meng Q, Du B and Zhou S 2026 Federated transfer learning-based model for remaining useful life prediction of rolling bearing across edge clients *Meas. Sci. Technol.* **37** 026007
[21] Bisoyi S, Rathi A K and Mahato S 2025 Fault diagnosis of rolling bearing failures using a multi-stage e-CNN-GRU-SAM network *Sci. Rep.* **15** 33102
[22] Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P and Shroff G 2016 LSTM-based encoder-decoder for multi-sensor anomaly detection (arXiv:1607.00148)
[23] Zhang D, Zhou S, Zheng Y and Xu X 2025 Review on application of machine vision-based intelligent algorithms in gear defect detection *Processes* **13** 3370
[24] Garcia J, Rios-Colque L, Peña A and Rojas L 2025 Condition monitoring and predictive maintenance in industrial equipment: an NLP-assisted review of signal processing, hybrid models and implementation challenges *Appl. Sci.* **15** 5465
[25] Chang X, Zhang Y, Liu M, Wang T and Wang H 2025 Proactive health management strategies for older adults with motoric cognitive risk syndrome: a scoping review *BMC Geriatr.* **26** 142
[26] Prasad A, Dantreliya C, Chande M, Chauhan V and Rai A 2023 An intelligent fault diagnosis framework based on piecewise aggregate approximation, statistical moments and sparse autoencoder *Proc. Inst. Mech. Eng. Pt. O J. Risk Reliab.* **237** 686–702
[27] Gupta M, Wadhvani R and Rasool A 2023 A real-time adaptive model for bearing fault classification and remaining useful life estimation using deep neural network *Knowl.-Based Syst.* **259** 110070
[28] Li J, Ding W, Mao W, Zhang J, Meng Z and Tong K 2024 Causal dilated convolution-based residual DenseNet with channel attention for RUL prediction of rolling bearings *Measurement* **235** 115012
[29] Cao W, Meng Z, Li J, Wu J and Fan F 2025 A remaining useful life prediction method for rolling bearing based on TCN-transformer *IEEE Trans. Instrum. Meas.* **74** 3501309
[30] Sankararaman S 2015 Significance, interpretation and quantification of uncertainty in prognostics and remaining useful life prediction *Mech. Syst. Signal Process.* **52–53** 228–47

[31] Zhang L, Lin J, Liu B, Zhang Z, Yan X and Wei M 2019 A review on deep learning applications in prognostics and health management *IEEE Access* **7** 162415–38
[32] Baraldi P, Cadini F, Mangili F and Zio E 2013 Model-based and data-driven prognostics under different available information *Probabilist. Eng. Mech.* **32** 66–79
[33] Lin Y H and Li G H 2022 A bayesian deep learning framework for RUL prediction incorporating uncertainty quantification and calibration *IEEE Trans. Ind. Informat.* **18** 7274–84
[34] Hu X, Tan L and Tang T 2024 M2BIST-SPNet: RUL prediction for railway signaling electromechanical devices *J. Supercomput.* **80** 16744–74
[35] Xuan Q L, Adhisantoso Y G, Munderloh M and Ostermann J 2023 Uncertainty-aware remaining useful life prediction for predictive maintenance using deep learning *Proc. CIRP* **118** 116–21
[36] Xu W and Zio E 2024 Uncertainty-aware prediction of remaining useful life in complex systems *Proc. Annual Conf. PHM Society 2024* vol 16 pp 1–4
[37] Zhang D, Dong J, Wang W, Bérenguer C and Zhao Z 2026 A multi-model ensemble and its portability for the prognostics of direct methanol fuel cells under different dynamic operating conditions *Reliab. Eng. Syst. Saf.* **270** 112179
[38] Yang C L, Meles T Y, Yilma A A and Teshome M M 2026 Uncertainty aware predictive maintenance using a hybrid transformer with Monte Carlo Dropout and conformal prediction *Ain Shams Eng. J.* **17** 103992
[39] Zhou J, Zhang H, Tang B, Zheng L and Wang Y 2026 Dynamic reliability informed adaptive task scheduling for multirobot manufacturing system *Adv. Eng. Inform.* **71** 104335
[40] Lei B, Wang Y, Li M, Liu S and Liu W 2026 Remaining useful life prediction with uncertainty quantification based on multi-scale attention-based physics-informed neural network *Eng. Res. Express* **8** 015521
[41] Song X, Liu J and Xia M 2023 Advanced vibration-based fault diagnosis and vibration control methods *Sensors* **23** 7704
[42] Zhang D, Stewart E, Ye J, Entezami M and Roberts C 2020 Roller bearing degradation assessment based on a deep MLP convolution neural network considering outlier regions *IEEE Trans. Instrum. Meas.* **69** 2996–3004
[43] Pan Y, Hong R, Chen J, Qin Z and Feng Y 2019 Incipient fault detection of wind turbine large-size slewing bearing based on circular domain *Measurement* **137** 130–42
[44] Rocco S and Claudio M 2013 Singular spectrum analysis and forecasting of failure time series *Reliab. Eng. Syst. Saf.* **114** 126–36
[45] Liu K, Law S S, Xia Y and Zhu X Q 2014 Singular spectrum analysis for enhancing the sensitivity in structural damage detection *J. Sound Vib.* **333** 392–417
[46] Khan M Z and Li D 2024 Dynamic logistic ensembles with recursive probability and automatic subset splitting for enhanced binary classification *2024 IEEE 15th Annual Ubiquitous Computing, Electronics and Mobile Communication Conf. (UEMCON)* pp 440–7
[47] García-León R A, Navarro-Barrera C J and Afanador-García N 2025 Colombian regulations in the seismic design of reinforced concrete buildings with portal frames: a comparative and bibliometric analysis *Buildings* **15** 4303
[48] Zhang J, Xiao T, Ji J, Zeng P and Cao Z J 2023 *Geotechnical Reliability Analysis: Theories, Methods and Algorithms* (Springer Nature)
[49] Matsumoto E Y and Del-Moral-Hernandez E 2016 Improving regression predictions using individual point reliability estimates based on critical error scenarios *Inf. Sci.* **374** 65–84
[50] Zhuang L, Wu X, Chow A H F, Ma W, Lam W H K and Wong S C 2025 Reliability-based journey time prediction via two-stream deep learning with multi-source data *J. Intell. Transp. Syst.* **29** 134–52
[51] Chang X, Guo J, Qin H, Huang J, Wang X and Ren P 2024 Single-objective and multi-objective flood interval forecasting considering interval fitting coefficients *Water Resour. Manage.* **38** 3953–72
[52] Taylor K E 2001 Summarizing multiple aspects of model performance in a single diagram *J. Geophys. Res. Atmos.* **106** 7183–92
[53] Wang B, Lei Y, Li N and Li N 2018 A hybrid prognostics approach for estimating remaining useful life of rolling element bearings *IEEE Trans. Reliab.* **69** 401–12
[54] Nectoux P, Gouriveau R, Medjaher K, Ramasso E, Chebel-Morello B, Zerhouni N and Varnier C 2012 PRONOSTIA: an experimental platform for bearings accelerated degradation tests *Proc. IEEE Int. Conf. Prognostics and Health Management (PHM 2012)*

Santosh Bisoyi received a BTech. degree in Civil Engineering from the College of Engineering and Technology, Bhubaneswar, Odisha, India, in 2019 and an MTech degree in Structural Engineering from the National Institute of Technology Agartala, Tripura, India, in 2021. He is currently working toward a PhD degree in Civil and Infrastructure Engineering at the Indian Institute of Technology Jodhpur, Rajasthan, India. His research interests include structural parameter identification, damage assessment, RUL, retrofitting, signal processing, machine learning, and DL.

Amit Kumar Rathi received a BE degree in Civil Engineering from Jai Narain Vyas University, India, in 2009, and an MTech degree from the Indian Institute of Technology Guwahati, India, in 2011. In 2019, he was awarded a PhD in Structural Engineering from the Indian Institute of Technology Guwahati, India. He is presently serving as an Assistant Professor in the Department of Civil and Infrastructure Engineering and Rishabh Centre Research and Innovation in Clean Energy (Secondary Affiliation) at the Indian Institute of Technology Jodhpur, Rajasthan, India. He has held faculty positions at the National Institute of Technology Sikkim, India and National Institute of Technology Calicut, India. His research focuses on reliability analysis and design, uncertainty quantification, stochastic modelling, structural dynamics, health monitoring, condition assessment and resilience.

Swarup Mahato received a BE degree in Civil Engineering from Bengal Engineering & Science University, Shibpur, in 2010, and MTech and PhD degrees in Structural Engineering from the Indian Institute of Technology Guwahati in 2013 and 2019, respectively. He has held postdoctoral and research positions at Université Gustave Eiffel, France; Kaunas University of Technology, Lithuania; and Aarhus University, Denmark, where his work focused on vibration-based monitoring, photogrammetric techniques, and digital twin–based sensing for structural systems. He is currently serving as an Assistant Professor in the Department of Civil Engineering at the Indian Institute of Engineering Science and Technology, Shibpur, India. His research broadly spans structural health monitoring, signal processing for condition assessment and predictive maintenance, time–frequency analysis, wireless sensing networks, vibration control, and bridge dynamics.