Ensemble of Data-Driven Prognostic Algorithms with Weight Optimization and K-Fold Cross Validation

Chao Hu[1], Byeng D. Youn[2,*] and Pingfeng Wang[3]

[1]*Department of Mechanical Engineering, University of Maryland, College Park, MD 20742, USA*
huchaost@umd.edu

[2]*School of Mechanical and Aerospace Engineering, Seoul National University, Seoul, 151-742, South Korea*
bdyoun@snu.ac.kr

[3]*Department of Industrial and Manufacturing Engineering, Wichita State University, Wichita, KS 67260, USA*
pingfeng.wang@wichita.edu

ABSTRACT

The traditional data-driven prognostic approach is to construct multiple candidate algorithms using a training data set, evaluate their respective performance using a testing data set, and select the one with the best performance while discarding all the others. This approach has three shortcomings: (i) the selected standalone algorithm may not be robust, i.e., it may be less accurate when the real data acquired after the deployment differs from the testing data; (ii) it wastes the resources for constructing the algorithms that are discarded in the deployment; (iii) it requires the testing data in addition to the training data, which increases the overall expenses for the algorithm selection. To overcome these drawbacks, this paper proposes an ensemble data-driven prognostic approach which combines multiple member algorithms with a weighted-sum formulation. Three weighting schemes, namely, the accuracy-based weighting, diversity-based weighting and optimization-based weighting, are proposed to determine the weights of member algorithms for data-driven prognostics. The k-fold cross validation (CV) is employed to estimate the prediction error required by the weighting schemes. Two case studies were employed to demonstrate the effectiveness of the proposed prognostic approach. The results suggest that the ensemble approach with any weighting scheme gives more accurate RUL predictions compared to any sole algorithm and that the optimization-based weighting scheme gives the best overall performance among the three weighting schemes.[†]

* Corresponding author.

1. INTRODUCTION

To support critical decision-making processes such as maintenance replacement and system design, activities of health monitoring and life prediction are of great importance to engineered systems composed of multiple components, complex joints, and various materials, such as distributed manufacturing facilities, electronic devices, advanced military systems and so on. Research on real-time diagnosis and prognosis which interprets data acquired by distributed sensor networks, and utilizes these data streams in making critical decisions provides significant advancements across a wide range of applications. Maintenance and life-cycle management is one of the beneficiary application areas because of the pervasive nature of design and maintenance activities throughout the manufacturing and service sectors. Maintenance and life-cycle management activities constitute a large portion of overall costs in many industries (Dekker, 1996). These costs are likely to increase due to the rising competition in today's global economy. For instance, in the manufacturing and service sectors, unexpected breakdowns can be prohibitively expensive since they immediately result in lost production, failed shipping schedules, no operational service, repair cost, and poor customer satisfaction. In order to reduce and possibly eliminate such problems, it is necessary to accurately assess current system health condition and precisely predict the remaining useful life (RUL) of operating components, subsystems, and systems.

In general, prognostics approaches can be categorized into model-based approaches (Luo, et al., 2008; Gebraeel & Pan, 2008; Gebraeel et al., 2009), data-driven approaches (Schwabacher, 2005; Wang et al., 2008; Zio & Di Maio, 2010; Coble & Hines, 2008; Heimes, 2008) and hybrid approaches (Kozlowski et al., 2001; Goebel et al., 2006; Saha et al., 2009). The application of general model-based prognostics

approaches relies on the understanding of system physics-of-failure and underlying system degradation models. Luo et al (Luo, et al., 2008) developed a model-based prognostic technique that relies on an accurate simulation model for system degradation prediction and applied this technique to a vehicle suspension system. Gebraeel presented a degradation modeling framework for RUL predictions of rolling element bearings under time-varying operational conditions (Gebraeel & Pan, 2008) or in the absence of prior degradation information (Gebraeel et al., 2009). As practical engineered systems generally consist of multiple components with multiple failure modes, understanding all potential physics-of-failures and their interactions for a complex system is almost impossible. With the advance of modern sensor systems as well as data storage and processing technologies, the data-driven approaches for system health prognostics, which are mainly based on the massive sensory data with less requirement of knowing inherent system failure mechanisms, have been widely used and become popular. A good review of data-driven prognostic approaches was given in (Schwabacher, 2005). Data-driven prognostic approaches generally require the sensory data fusion and feature extraction, statistical pattern recognition, and for the life prediction, the interpolation (Wang et al., 2008; Zio & Di Maio, 2010), extrapolation (Coble & Hines, 2008), or machine learning (Heimes, 2008) and so on. Hybrid approaches attempt to take advantage of the strength from data-driven approaches as well as model-based approaches by fusing the information from both approaches. Garga et al. (Kozlowski et al., 2001) described a data fusion approach where domain knowledge and predictor performance are used to determine weights for different state-of-charge predictors. Goebel et al. (Goebel et al., 2006) employed a Dempster-Shafer regression to fuse a physics-based model and an experience-based model for prognostics. Saha et al. (Saha et al., 2009) combined the offline relevance vector machine (RVM) with the online particle filter for battery prognostics. Similar to model-based approaches, the application of hybrid approaches is limited to the cases where sufficient knowledge on system physics-of-failures is available.

Implicit relationship between the RUL and the sensory signals makes it difficult to know which prognostic algorithm performs best in a specific application. Furthermore, there are many factors that affect the prediction accuracy and robustness, such as (i) dependency of the algorithm's accuracy on the number of units in a training data set, (ii) significant variability in manufacturing conditions and large uncertainties in environmental and operational conditions, (iii) the amount of effective sensory signals for RUL predictions, and (iv) the form of degradation trend (e.g., linear, nonlinear, noisy, smooth). Therefore, no single prognostic algorithm works well for all possible situations. Instead of using an individual prognostic algorithm, it would be beneficial to combine multiple algorithms to form a hybrid algorithm. Combining different approximate algorithms into an ensemble has found its applications in a wide variety of research fields, such as the development of committees of neural networks (Perrone & Cooper, 1993; Bishop, 2005), the metamodeling for the design of modern engineered systems (Goel et al., 2007; Acar & Rais-Rohani, 2009), the discovery of regulatory motifs in bioinformatics (Hu et al., 2006), the detection of traffic incidents (Chen et al., 2009), and the development of ensemble Kalman filters (Evensen, 2003). However, the utilization of the ensemble approach for the data-driven prognostics is still in infancy. Most data-driven prognostic practices select a single algorithm with the best accuracy from the algorithm pool while discarding the others. This approach not only wastes the resource devoted to developing different algorithms, but also suffers from the lack of robustness.

Estimating the accuracy of a prognostic algorithm is important not only for evaluating its prediction accuracy but also for choosing the best algorithm from a given set (model selection), or combining algorithms. Many data-driven approaches (Schwabacher, 2005; Wang et al., 2008; Zio & Di Maio, 2010) use the so-called holdout method, which divides the original run-to-failure data set into two mutually exclusive subsets called a training set and a testing set, or holdout set. The holdout method is straightforward and computationally efficient. However, it often produces a large variance of the resulting estimate and requires the testing data set which increases the overall expenses for the algorithm selection.

To overcome the above shortcomings, this study proposes an ensemble approach that employs the *k*-fold cross validation (CV) to estimate the accuracy of a given ensemble and proposes three weighting schemes to determine the weight values. Assumptions for this study are listed below:

(1) Multiple run-to-failure data are available, either from the computer simulation or field testing.
(2) A single failure mode is considered, i.e., the RUL prediction is exclusively for this failure mode.
(3) The underlying physics of the system fault propagation is not comprehensive or it is too expensive to derive a reliable physical damage model for a complex engineered system. Both cases entail the use of the data-driven prognostics.

These assumptions define the application domain of this work, i.e., data-driven prognostics. The rest of the paper is organized as follows. Section 2 presents the proposed ensemble approach with the *k*-fold CV and three weight schemes. Applications of the proposed

methodology are presented in Section 3 and the conclusion of this work is given in Section 4.

2. ENSEMBLE OF PROGNOSTIC ALGORITHMS

It is essential to propose a robust prognostic solution that accurately predicts the RUL using data features extracted from multi-dimensional sensory degradation signals. For building such a unified structural health prognostic framework, this paper proposes (i) a weighted-sum formulation for an ensemble of prognostic algorithms, (ii) *k*-fold cross validation (CV) to evaluate the error metric associated with a candidate ensemble model; and (iii) three weighting scheme to determine the weight values for the member algorithms. This section is organized as follows. Section 2.1 presents the basic weighted-sum formulation for the RUL prediction. Section 2.2 describes the background of the *k*-fold CV and how it can be applied for estimating the accuracy of a prognostic algorithm. Section 2.3 describes the three proposed weighting schemes. The overall procedure of the ensemble approach is described in Section 2.4.

2.1 Weighted-Sum Formulation

The weighted-sum and voting formulations are most often used for the algorithm ensemble. Since the RUL is not binary, the weighted-sum approach which combines RUL predictions from all member algorithms is more approximate than the voting formulation which retains only one RUL prediction while discarding all the others. A simple average of RUL predictions obtained using the member algorithms is acceptable only when the member algorithms provide the same level of accuracy. However, it is more likely that an algorithm tends to be more accurate than others. It is ideal to assign a greater weight to a member algorithm with higher prediction accuracy in order to enhance its prediction accuracy and robustness.

Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2,\ldots, \mathbf{y}_N\}$ be a data set consisting of multi-dimensional sensory signals (e.g., acceleration, strain, pressure) from N different run-to-failure units. An ensemble of prognostic member algorithms for RUL prediction can be expressed in a weighted-sum formulation as

$$\hat{L} = \sum_{j=1}^{M} w_j \hat{L}_j \left(\mathbf{y}_t, \mathbf{Y}\right) \tag{1}$$

where $\hat{L}$ denotes the ensemble predicted RUL for the testing data set $\mathbf{y}_t$; M denotes the number of algorithm members in the ensemble; w_j denotes the weight assigned to the j^{th} prognostic algorithm; $\hat{L}_j$ ($\mathbf{y}_t$, $\mathbf{Y}$) denotes the predicted RUL by the j^{th} prognostic member algorithm trained with the data set $\mathbf{Y}$.

2.2 *K*-Fold Cross Validation

The *k*-fold cross validation is used in the offline process to evaluate the accuracy of member algorithms and a given ensemble. It randomly divides the original data set $\mathbf{Y}$ into k mutually exclusive subsets (or folds) $\mathbf{Y}_1$, $\mathbf{Y}_2,\ldots,$ $\mathbf{Y}_k$ having an approximately equal size (Kohavi, 1995). Of the k subsets, one is used as the test set and the other $k-1$ subsets are put together as a training set. The CV process is performed k times, with each of the k subsets used exactly once as the test set. Let $\mathbf{I}_m = \{i: \mathbf{y}_i \in \mathbf{Y}_m\}$, $m = 1, 2,\ldots, k$ denote the index set of the run-to-failure units whose sensory signals construct the subset $\mathbf{Y}_m$. Then the CV error is computed as the average error over all k trials and can be expressed as

$$\varepsilon_{CV} = \frac{1}{N}\sum_{m=1}^{k}\sum_{i\in\mathbf{I}_m} S\left(\hat{L}\left(w_j, \hat{L}_j\left(\mathbf{y}_i, \mathbf{Y}\setminus\mathbf{Y}_m\right)\right), L_i^T\right) \tag{2}$$

where $S(\bullet)$ is a predefined evaluation metric that measures the accuracy of the ensemble-predicted RUL; N denotes the number of different run-to-failure units for CV; L_i^T denotes the true RUL of the i^{th} unit. The above formula indicates that all units in the data set are used for both training and testing, and each unit is used for testing exactly once and for training $k-1$ times. Thus, the variance of the resulting estimate is likely to be reduced compared to the traditional holdout approach, resulting in superior performance when employing a small data set. It is important to note that the disadvantage of the *k*-fold CV against the holdout method is greater computational expense because the training process has to be executed k times. As a commonly used setting for CV, a 10-fold CV is employed in this study.

2.3 Weighting Schemes

This section will introduce three schemes to determine the weights of member algorithms: the accuracy-based weighting, diversity-based weighting and optimization-based weighting.

2.3.1 Accuracy-based Weighting

The prediction accuracy of the j^{th} member algorithm is quantified by its CV error, expressed as

$$\varepsilon_{CV}^{j} = \frac{1}{N}\sum_{m=1}^{k}\sum_{i\in\mathbf{I}_m} S\left(\hat{L}_j\left(\mathbf{y}_i, \mathbf{Y}\setminus\mathbf{Y}_m\right), L_i^T\right) \tag{3}$$

The weight w_j of the j^{th} member algorithm can then be defined as the normalization of the corresponding inverse CV error as

$$w_j = \frac{\left(\varepsilon_{CV}^{j}\right)^{-1}}{\sum_{i=1}^{M}\left(\varepsilon_{CV}^{i}\right)^{-1}} \tag{4}$$

This definition indicates that a larger weight is assigned to a member algorithm with higher prediction accuracy.

Thus, a member algorithm with better prediction accuracy has a larger influence on the ensemble prediction. This weighting scheme relies exclusively on the prediction accuracy to determine the weights of member algorithms.

2.3.2 Diversity-based Weighting

The weight formulation in Eq. (4) relies exclusively on the prediction accuracy to determine the weights. However, the prediction accuracy of member algorithms is not the only factor that affects the ensemble performance. The prediction diversity, which measures the extent to which the predictions by a member algorithm are distinguishable from those by the others, also has a significant effect on the ensemble performance, especially on the robustness. More specifically, a larger weight should generally be assigned to a member algorithm with higher prediction diversity because of its larger potential to enhance the ensemble robustness.

We begin by formulating an N-dimensional error vector consisting of absolute RUL prediction errors by the j^{th} member algorithm as

$$\mathbf{e}_j = \left(\hat{L}_j \left(\mathbf{y}_1, \mathbf{Y} \setminus \mathbf{Y}_1 \right) - L_1^T, ..., \hat{L}_j \left(\mathbf{y}_N, \mathbf{Y} \setminus \mathbf{Y}_m \right) - L_N^T \right)^{\mathrm{T}} \quad (5)$$

Repeatedly computing the error vectors for all M member algorithms gives M error vectors $\mathbf{e}_1$, $\mathbf{e}_2$,…, $\mathbf{e}_M$. The prediction diversity of the j^{th} member algorithm can then be computed as the sum of Euclidean distances between the error vector $\mathbf{e}_j$ and all the other error vectors, given by

$$D_j = \sum_{i=1; i \neq j}^{k} \left\| \mathbf{e}_j - \mathbf{e}_i \right\| \quad (6)$$

The prediction diversity measures the extent to which the predictions by a member algorithm are distinguishable from those by any other. Based on the defined prediction diversity, the normalized weight w_j of the j^{th} member algorithm can then be calculated as

$$w_j = \frac{D_j}{\sum_{i=1}^{M} D_i} \quad (7)$$

This definition suggests that a member algorithm with higher prediction diversity will be given a larger weight and thus contributes more to the ensemble predicted RUL. For example, if, among all the member algorithms, one algorithm consistently gives early RUL predictions while any of the others late RUL predictions, the former will likely be given a larger weight than the latter. It is also noted that the weight formulation in Eq. (7) considers the prediction diversity as the only criterion for the weight determination.

2.3.3 Optimization-based Weighting

Neither the accuracy-based nor diversity-based weighting scheme takes into account both the prediction accuracy and diversity in the weight calculation. Thus, the two schemes cannot produce an ensemble algorithm to achieve both high prediction accuracy and robustness. In what follows, an optimization-based weighting scheme is proposed to maximize the accuracy and robustness of data-driven prognostics by adaptively synthesizing the prediction accuracy and diversity of each member algorithm.

In the optimization-based weighting scheme, the weights in Eq. (1) can be obtained by solving an optimization problem of the following form

$$\begin{aligned} &\text{Minimize } \varepsilon_{CV} \left(\hat{L} \left(w_j, \hat{L}_j \left(\mathbf{y}^i \right) \right), L_T^i, i = 1, ..., N \right) \\ &\text{Subject to } \sum_{j=1}^{M} w_j = 1 \end{aligned} \quad (8)$$

After the prediction of RULs using the M member algorithms through the 10-fold CV, the above optimization problem can be readily solved with almost negligible computational effort since the weight optimization process does not require the execution of member algorithms. Thus, the overall computational cost mainly comes from the training and testing in the CV process. We expect that, by solving the optimization problem in Eq. (8), the resulting ensemble of algorithms will outperform any of the ensemble's individual member algorithms in terms of both accuracy and robustness.

2.3.4 Overall Procedure

Figure 1 shows the overall procedure of the proposed ensemble approach with the k-fold CV and three weighting schemes. This data-driven prognostic approach is composed of the offline and online processes. In the offline process, the offline training/testing process with the k-fold CV is employed to compute the CV error of an ensemble formulation; the weights of member algorithms are determined using the accuracy-based weighting, diversity-based weighting and optimization-based weighting. The online prediction process combines the RUL predictions from all member algorithms to form an ensemble RUL prediction using the weights obtained from the offline process. This process enables the continuous update of the health information and prognostic results in real-time with new sensory signals. STEPS 2-4 can be repeated to use new training sensory signals and to update the weights and RUL predictions. Since the computationally expensive training process with multiple algorithms is done offline and the online prediction process with multiple algorithms requires a small amount of computational effort, the ensemble approach raises little concerns in the computational

complexity. Indeed, in many engineered systems, the prognostic accuracy is treated as of much more importance compared to the computational complexity since the occurrence of a catastrophic system failure causes much more loss than the increase of the computational efforts. Therefore, in cases where the ensemble approach achieves significant improvement in the prediction accuracy compared to any sole member algorithm, we should always prefer the use of the former.

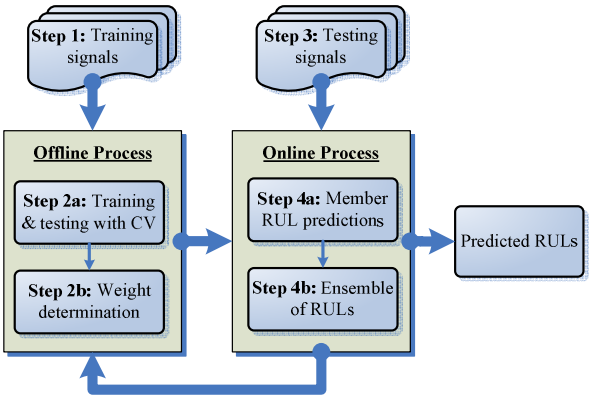


Figure 1: A flowchart of the ensemble approach

3. CASE STUDIES

In this section, the proposed ensemble of data-driven prognostic algorithms is demonstrated with two case studies: (i) 2008 IEEE PHM challenge problem and (ii) power transformer problem. In each case study, the ensemble approach combines RUL predictions from five data-driven prognostic algorithms, namely, a similarity-based interpolation (SBI) approach with RVM as the regression technique (RVM-SBI) (Wang et al., 2008; Tipping, 2001), SBI with SVM (SVM-SBI) (Wang et al., 2008; Smola & Schölkopf, 2004), SBI with the least-square exponential fitting (Exp-SBI) (Wang et al., 2008), a Bayesian linear regression with the least-square quadratic fitting (Quad-BLR) (Coble & Hines, 2008), and a recurrent neural network (RNN) approach (RNN) (Heimes, 2008; Cernansky et al., 2007).

3.1 2008 IEEE PHM Challenge Problem

3.1.1 Description of Data Set

The data set provided by the 2008 IEEE PHM Challenge problem consists of multivariate time series signals that are collected from an engine dynamic simulation process. Each time series signal comes from a different degradation instance of the dynamic simulation of the same engine system (Saxena & Goebel, 2008). The data for each cycle of each unit include the unit ID, cycle index, 3 values for an operational setting and 21 values for 21 sensor measurements. The sensor data were contaminated with measurement noise and different engine units start with different initial health conditions and manufacturing variations which are unknown. Three operational settings have a substantial effect on engine degradation behaviors and result in six different operation regimes. The 21 sensory signals were obtained from six different operation regimes. The whole data set was divided into training and testing subsets, each of which consists of 218 engine units. In the training data set, the fault growth in a unit was allowed until the occurrence of a system failure when one or more limits for safe operation have been reached. In the testing data set, the time series signals were pruned some time prior to the occurrence of a system failure. The objective of the problem is to predict the number of remaining operational cycles before failure in the testing data set.

3.1.2 Implementation of Ensemble Approach

For the CV process, the training data set with 218 units were divided to 10 data subsets with a similar size. Each data subset was used for both training and testing and, more specifically, 9 times for training and once for testing. The training data subsets contain complete degradation information while the testing data subsets carry only partial degradation information. The latter were generated by truncating the original data subsets after pre-assigned RULs. The RUL pre-assigned to each unit in a testing data subset was randomly generated from a uniform distribution between zero and its half life. This range in the uniform distribution was selected based on the following two criteria: (i) the pre-assigned RULs should be small enough to allow the occurrence of substantial degradation; and (ii) the variation of the pre-assigned RULs should be large enough to test the robustness of algorithms.

Following the previous work (Wang et al., 2008), this study selected 7 sensory signals (2, 3, 4, 7, 11, 12 and 15) among the 21 sensory signals for the use in the member algorithms: RVM-SBI, SVM-SBI, Exp-SBI and Quad-BLR. For the VHI construction, the system healthy matrix $\mathbf{Q}_0$ was created with the sensory data in a system healthy condition, $L > 300$, while the system failure matrix $\mathbf{Q}_1$ with those in a system failure condition, $0 \le L \le 4$. The RVM employed a linear spline kernel function with the initial most probable hyper-parameter vector for kernel weights $\boldsymbol{\alpha}_m = [1\times10^4,\ldots, 1\times10^4]$ and the initial most probable noise variance $\sigma_m^2 = 1\times10^{-4}$. In the SVM, a Gaussian kernel function is used with the parameter settings as: the regularization parameter $C = 10$ and the parameter of the ε-insensitive loss function $\varepsilon = 0.10$. The implementation details of RNN can be found in (Heimes, 2008). In the RNN architecture, the numbers of the input, recurrent and output units are $|I| = 22$, $|R| = 8$ and $|O| = 1$.

The evaluation metric considered for this example employed an asymmetric score function around the true RUL such that heavier penalties are placed on late predictions (Saxena & Goebel, 2008). The score evaluation metric *S* can be expressed as

$$S\left(\hat{L}_i, L_i^T\right) = \begin{cases} \exp\left(-d_i/13\right)-1, & d_i < 0 \\ \exp\left(d_i/10\right)-1, & d_i \geq 0 \end{cases} \quad (9)$$

where $d_i = \hat{L}_i - L_i^T$

where $\hat{L}_i$ and L^T denote the predicted and true RUL of the i^{th} unit, respectively. This score function was used to compute the CV error ε_{CV} using Eq. (2) for the accuracy- and optimization-based weighting schemes. In this study the weight optimization problem in Eq. (8) was solved using a sequential quadratic optimization (SQP) method which is a gradient-based optimization technique.

3.1.3 Results of Ensemble Approach

The five selected member algorithms are RVM-SBI (RS), SVM-SBI (SS), Exp-SBI (ES), Quad-BLR (QB) and RNN (RN). The three weighting schemes are the accuracy-based weighting (AW), diversity-based weighting (DW) and optimization-based weighting (OW). Table 1 summarizes the weight optimization results of the ensemble approaches as well as compares the CV and validation errors of the individual and ensemble approaches. It is observed that the ensemble approaches with all three weighting schemes outperforms any of the individual member algorithm in terms of the CV error and that the one with the optimization-based weighting achieves the smallest CV error of 4.8387 on the training data set, a 38.62% improvement over the best individual member algorithm, ES, whose CV error is 7.8834. As expected, the accuracy-based weighting scheme yields better prediction accuracy than the diversity-based weighting. This can be attributed to the fact that the former assigns larger weights to member algorithms with better prediction accuracy while the latter does not consider the prediction accuracy in the weight determination. To test the robustness of the ensemble approaches, the testing data set with 218 units were employed to compute the validation errors. Note that the testing data set is different from the training data set that was used to determine the weights in the ensemble approach. It is remarkable that the ensemble approaches again outperform the individual member algorithms and that the one with the diversity-based weighting performs best, with a 34.7% improvement over the best individual member algorithm, SS. This suggests that the diversity-based weighting, compared to the accuracy-based weighting, provides a more robust ensemble of the member algorithms. It is noted that the optimization-based weighting scheme still achieves a comparable validation error to that of the diversity-based weighting scheme.

Under the optimization-based weighting scheme, the RUL predictions by two individual algorithms, ES and QB, with the largest weights and the ensemble approach are plotted for 218 training and testing units in Figure 2. The units are sorted by the RULs in an ascending order. It is seen that ES tends to give consistently early RUL predictions while QB tends to provide consistently late RUL predictions. We also observed that QB exhibited a significant "unbalanced" prediction feature while ES possesses this feature to a relatively small extent. Such a feature of QB is due to the fact that the Bayesian linear regression with a quadratic projection scheme, as its online prediction technique, cannot fully capture the exponential degradation trend and often over-project the degradation trend. In contrast, the ensemble approach gives RUL predictions closer to the true values while eliminating many outliers produced by the two individual algorithms. The optimization-based weighting scheme provides better performance since the scheme employs an optimum ensemble formulation.

Table 1: Weighting results, CV and validation errors for 2008 PHM challenge problem

	RS	SS	ES	QB	RN	RS-SS-ES-QB-RN		
						AW	DW	OW
Weight by AW	0.3063	0.3029	0.3137	0.0151	0.0620	---	---	---
Weight by DW	0.1478	0.1488	0.1488	0.3354	0.2191	---	---	---
Weight by OW	0.0000	0.0470	0.7462	0.2068	0.0000	---	---	---
CV error	8.0743	8.1646	7.8834	163.3376	39.8583	6.9159	7.0852	**4.8387**
Validation error	10.2393	9.3907	10.4710	247.0079	20.1499	8.5544	**6.1280**	6.1955

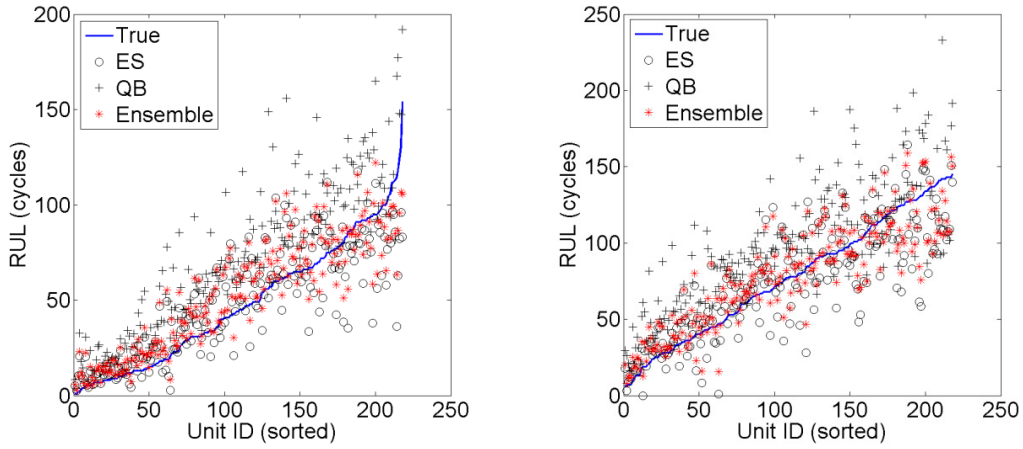


Figure 2: RUL predictions of training units (a) and testing units (b) for 2008 PHM challenge problem (optimization-based weighting)

3.1.4 Comparison of Different Combinations of Member Algorithms

Out of the five member algorithms, 31 different combinations can be chosen to formulate an ensemble approach. It would be interesting to study how a choice of combination affects the performance of an ensemble approach. Table 2 summarizes the CV errors for ensemble approaches with all possible combinations of the member algorithms under the optimization-based weighting scheme. Three important remarks can be derived from the results. First of all, it is observed that the ES, as the individual member algorithm with the best performance, always serves as a member algorithm of the best ensemble approach. We also observe that the ES, when involved in the ensemble approach, always had a larger weight than any other. It indicates that the best member algorithm exhibits good cooperative performance which can be identified by the optimization-based weighting scheme. Secondly, the QB, which gives the worst individual performance, was surprisingly selected as an important member of the best ensemble approach. These results, though counterintuitive, suggest that the ensemble approach can adaptively synthesize the prediction ability and diversity of each individual algorithm to enhance the accuracy and robustness of RUL predictions. Indeed, the QB is prone to give late RUL predictions as shown in Figure 2 and thus possesses higher prediction diversity. Thirdly, both the mean and standard deviation of CV errors decrease as the number of member algorithms increases. The mean and standard deviation of CV errors of ensemble approaches with a single member algorithm are 45.4636 and 67.3188, respectively, and they monotonically decrease to 5.1896 and 0.7440, respectively, by the ensemble approach with four member algorithms. Thus it would be beneficial to have more member algorithms to enhance the prediction accuracy and reduce the uncertainty of this accuracy.

3.2 Power Transformer Problem

3.2.1 Model Description

Power transformers are among the most expensive elements of high-voltage power systems. The monitoring of power transformers enables the transition from the traditional time-based maintenance to the condition-based maintenance, resulting in significant reductions in maintenance costs (Leibfried, 1998). Since it is very difficult, if not impossible, to obtain direct measurements of the health condition of transformers, indirect measurements are most often used to diagnose the health condition and predict the RUL of transformers (Rivera et al., 2000).

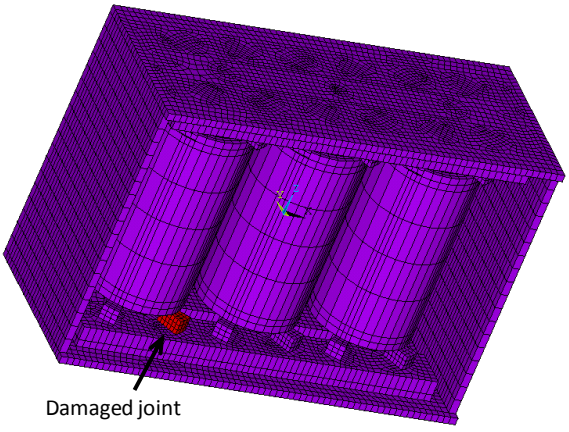


Figure 3: A power transformer FE model

Table 2: Comparison of CV errors of different combinations of member algorithms (optimization-based weighting)

Combination	CV error	Combination	CV error	Combination	CV error
RS	8.0743	RS-SS	8.0769	RS-SS-ES	7.8834
SS	8.1646	RS-ES	7.8834	RS-SS-QB	4.9123
ES	**7.8834**	RS-QB	4.9162	RS-SS-RN	6.7983
QB	163.3376	RS-RN	6.8002	RS-ES-QB	4.8391
RN	39.8583	SS-ES	7.8834	RS-ES-RN	6.5194
Mean	*45.4636*				
Std[a]	*67.3188*				
RS-SS-ES-QB	**4.8387**	SS-QB	4.9362	RS-QB-RN	4.9162
RS-SS-ES-RN	6.5194	SS-RN	6.8376	**SS-ES-QB**	**4.8387**
RS-SS-QB-RN	4.9123	**ES-QB**	**4.8391**	SS-ES-RN	6.5194
RS-ES-QB-RN	4.8391	ES-RN	6.5194	SS-QB-RN	4.9362
SS-ES-QB-RN	**4.8387**	QB-RN	17.5868	ES-QB-RN	4.8391
Mean	*5.1896*	*Mean*	*7.6279*	*Mean*	*5.7002*
Std	*0.7440*	*Std*	*3.7182*	*Std*	*1.1234*
RS-SS-ES-QB-RN	4.8387				

[a] *Standard deviation*

This case study aims to use sensory measurements of the transformer vibration responses induced by the magnetic field loading to predict the RUL of a winding support joint against a mechanical failure (loosening). The finite element (FE) model of a power transformer was created in ANSYS 10 as shown in Figure 3, where one exterior wall is concealed to make the interior structure visible. The transformer is fixed at the bottom surface and a vibration load with the frequency of 120 Hz is applied to the magnetic core. The three windings have a total number of twelve support joints, with each having four support joints. The joint loosening was simulated by reducing the stiffness of the joint. The random parameters considered in this study are listed in Table 3, which includes the material properties of support joints and windings as well as the geometries of the transformer. The uncertainties in vibration responses propagated from these uncertain parameters will be accounted for when generating prognostic data.

Table 3: Random geometries and material properties for power transformer problem

Component	Physical meaning	Distri. type	Mean	Std
x_1	Wall Thickness	Normal	3	0.015
x_2	Angular width of support joints	Normal	15	0.075
x_3	Height of support joints	Normal	6	0.03
x_4	Young's modulus of support joint	Normal	2E+12	1E+10
x_5	Young's modulus of winding	Normal	1.28E+12	6E+8
x_6	Poisson ratio of joints	Normal	0.27	0.0027
x_7	Poisson ratio of winding	Normal	0.34	0.0034
x_8	Density of joints	Normal	7.85	0.000785
x_9	Density of windings	Normal	8.96	0.0896

3.2.2 Prognostic Data Generation

The failure mode considered in this study is the loosening of a winding support joint (see Figure 3) induced by the magnetic core vibration. The failure criterion is defined as a 99% stiffness reduction of the joint. To model the trajectory of change in stiffness over time, this study uses a damage propagation model with an exponential form as (Saxena & Goebel, 2008)

$$E(t) = E_0 + b_E\left(1 - \exp\left(a_E t\right)\right) \tag{10}$$

where E_0 is the initial Young's modulus of the joint; a_E and b_E are the model parameters; t is the cycle time. The initial Young's modulus E_0 follows the same normal distribution with x_4 (see Table 3). The model parameters a_E and b_E are independent and normally distributed with means 0.002 and 4E+12, each of which has a 10% coefficient of variation.

Since data-driven prognostic approaches require a large amount of prognostic data, it is computationally intolerable, if not impossible, to simply run the simulation to generate every data point. To overcome this difficulty, this study employed the univariate decomposition method that only uses a certain number of univariate sample points to accurate construct the response surface for a general multivariate response function while achieving good accuracy (Xu & Rahman, 2005).

This study selected 5 sensors from the optimally designed sensor network consisting of 9 sensors and thus requires the construction of 5 responses surfaces. The data generation process involves four sequentially executed procedures: (i) four univariate sample points were obtained from the harmonic analysis to construct response surfaces, along the damage propagation path, that approximate the strain components at five sensor locations as functions of random variables detailed in Table 3; (ii) 400 randomly generated samples of E_0, a_E and b_E were used in conjunction with Eq. (10) to produce 400 damage propagation paths were produced with randomly generated, of which 200 paths were assigned to the training units and the rest to the testing units; (iii) the constructed response surfaces were used to interpolate the strain components at five sensor locations for a given set of randomly generated geometries and material properties and damage propagation path, and repeatedly executing this process for 400 times gives the training data set with 200 training units and the testing data set with 200 testing units; (iv) measurement noise following a zero mean normal distribution was added to both the training and testing data sets to finalize the data generation. The cubic spline was used as the numerical scheme for the response surface construction and interpolation. Simulated measurements by sensor 1 are plotted against the adjusted cycle index, defined as the subtraction of the cycle-to-failure from the actual operational cycle, in Figure 4, for all 200 training units in the training data set.

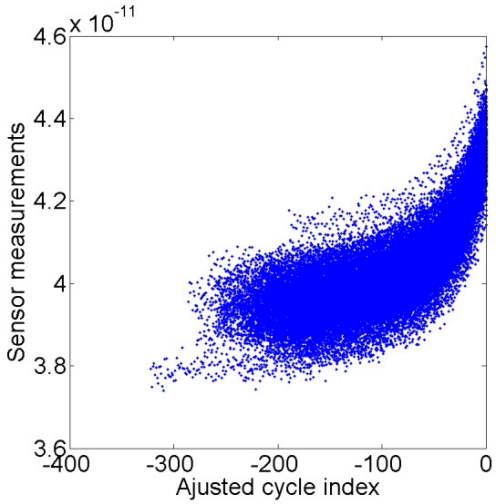


Figure 4: Simulated measurements by sensors 1

3.2.3 Implementation of Ensemble Approach

The training data set with 200 units were equally and randomly divided to 10 subsets. Similar to the first example, when used for the testing in CV, each unit in a subset was assigned with a randomly generated RUL from a uniform distribution between zero and its half life. All the five member algorithms used the same parameter settings with those detailed in Section 3.1.2. The score function in Eq. (9) was again used to compute the CV error ε_{CV} for the accuracy- and optimization-based weighting schemes.

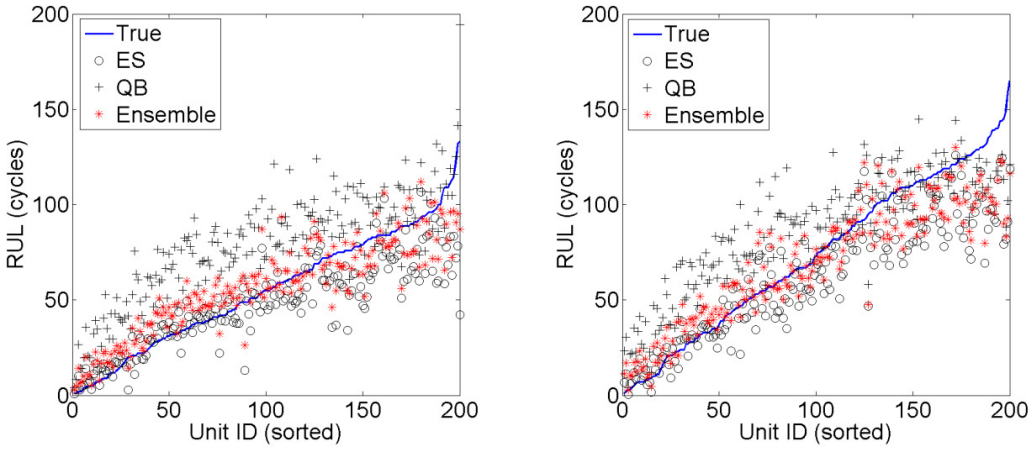
3.2.4 Results of Ensemble Approach

Table 4 summarizes the weight optimization results of the ensemble approaches as well as compares the CV and validation errors of the individual and ensemble approaches. Compared to the first example, similar results can be observed: (i) the ensemble approaches with all three weighting schemes yield smaller CV error than any of the individual member algorithm and the one with the optimization-based weighting gives the smallest CV error of 2.7258 on the training data set, a 66.48% improvement over the best individual member algorithm, RN, whose CV error is 8.1323; (ii) the accuracy-based weighting scheme yields a comparable CV error to that of the diversity-based weighting; (iii) the optimization-based weighting scheme achieves a validation error of 5.6138, which is comparable to the smallest validation error of 5.6119 by the diversity-based weighting scheme.

Under the optimization-based weighting scheme, the RUL predictions by two individual algorithms, ES and QB, with the largest weights and the ensemble approach are plotted for 218 training and testing units in Figure 5. It can be observed that ES and QB are prone to produce late and early RUL predictions, respectively, while the ensemble approach gives RUL predictions closer to the true values with a much smaller number of outliers.

Table 4: Weighting results, CV and validation errors for power transformer problem

	RS	SS	ES	QB	RN	RS-SS-ES-QB-RN		
						AW	DW	OW
Weight by AW	0.2128	0.2265	0.2343	0.0677	0.2588	---	---	---
Weight by DW	0.1488	0.1486	0.1688	0.3290	0.2048	---	---	---
Weight by OW	0.0000	0.0000	0.6303	0.2336	0.1361	---	---	---
CV error	9.8922	9.2945	8.9849	31.0891	8.1323	3.4874	3.4124	**2.7258**
Validation error	6.5737	6.8847	7.8251	20.0356	15.2265	5.7825	**5.6119**	5.6138



(a) (b)

Figure 5: RUL predictions of training units (a) and testing units (b) for power transformer problem (optimization-based weighting)

4. CONCLUSION

This paper proposed a novel ensemble approach for the data-driven prognostics, which employed the *k*-fold cross validation and three weighting schemes. By combining the predictions of all member algorithms, the ensemble approach achieves better accuracy in RUL predictions compared to any sole member algorithm. Furthermore, the ensemble approach has an inherent flexibility to incorporate any advanced prognostic algorithm that will be newly developed. To the best of our knowledge, this is the first study of an ensemble approach with three weighting scheme for the data-driven prognostics. Since the computationally expensive training process is done offline and the online prediction process requires a small amount of computational effort, the ensemble approach raises little concerns in the computational feasibility. Two engineering case studies (2008 PHM challenge problem and the power transformer prognostics problem) demonstrated the superb performance of the proposed ensemble approach for the data-driven prognostics. Among the three weighting scheme, the optimization-based weighting scheme showed the capability of adaptively synthesizing the prediction accuracy and diversity of each member algorithm to enhance the accuracy of RUL predictions. Considering the enhanced accuracy in RUL predictions and flexibility in algorithm incorporations, the proposed ensemble approach is a promising methodology for the data-driven prognostics.

ACKNOWLEDGMENT

The authors would like to acknowledge that this research is partially supported by US National Science Foundation (NSF) under Grant No. GOALI-0729424, U.S. Army TARDEC by the STAS contract (TCN-05122), and by General Motors under Grant No. TCS02723.

NOMENCLATURE

a_E	1st degradation model parameter
b_E	2nd degradation model parameter
C	regularization parameter
d_i	difference between i^{th} predicted and true RULs
D_j	diversity of the j^{th} member algorithm
$\mathbf{e}_j$	error vector the j^{th} member algorithm
E	Young's modulus
E_0	initial Young's modulus
$\|I\|$	numbers of the input units
$\mathbf{I}_m$	index set the m^{th} subset
k	number of folds in cross validation
$\hat{L}$	ensemble predicted RUL
L_i^T	true RUL of the i^{th} unit
$\hat{L}_j$	predicted RUL by the j^{th} prognostic algorithm
N	number of run-to-failure units
$\|O\|$	numbers of the output units
$\mathbf{Q}_0$	system healthy matrix
$\mathbf{Q}_1$	system failure matrix
$\|R\|$	numbers of the recurrent units
$S(\bullet)$	evaluation metric
t	cycle time
$\mathbf{Y}$	prognostic data set
w_j	weight of the j^{th} prognostic algorithm
$\boldsymbol{\alpha}_m$	initial most probable hyper-parameter vector
σ_m^2	initial most probable noise variance
ε	parameter of the ε-insensitive loss function
ε_{CV}	cross validation error
CV	cross validation
RUL	remaining useful life
RVM	relevance vector machine
SBI	similarity-based interpolation
SVM	support vector machine

REFERENCES

Dekker, R. (1996). Applications of maintenance optimization models: a review and analysis, *Reliability Engineering and System Safety*, vol. 51, no. 3, pp. 229–240.

Luo, J., Pattipati, K.R., Qiao, L. & Chigusa, S. (2008). Model-based prognostic techniques applied to a suspension system, *IEEE Transactions on Systems, Man and Cybernetics, Part A,* vol. 38, no. 5, pp. 1156–1168.

Gebraeel, N. & Pan, J. (2008). Prognostic degradation models for computing and updating residual life distributions in a time-varying environment, *IEEE Transactions on Reliability*, vol. 57, no. 4, pp. 539–550.

Gebraeel, N., Elwany, A. & Pan J. (2009). Residual life predictions in the absence of prior degradation knowledge, *IEEE Transactions on Reliability*, vol. 58, no. 1, pp. 106–117.

Schwabacher, M. (2005). A survey of data-driven prognostics, *Proceedings of AIAA Infotech@Aerospace Conference*, Arlington, VA.

Wang, T., Yu, J., Siegel, D. & Lee, J. (2008). A similarity-based prognostics approach for remaining useful life estimation of engineered systems, *IEEE, International Conference on Prognostics and Health Management*, Denver, CO, Oct 6-9.

Zio, E. & Di Maio, F. (2010). A data-driven fuzzy approach for predicting the remaining useful life in dynamic failure scenarios of a nuclear power plant, *Reliability Engineering and System Safety*, vol. 95, no. 1, pp. 49–57.

Coble, J.B. & Hines, J.W. (2008). Prognostic algorithm categorization with PHM challenge application, *IEEE, International Conference on Prognostics and Health Management*, Denver, CO, Oct 6-9.

Heimes, F.O. (2008) Recurrent neural networks for remaining useful life estimation, *IEEE, International Conference on Prognostics and Health Management*, Denver, CO, Oct 6-9.

Kozlowski, J.D., Watson, M.J., Byington, C.S., Gargam A.K. & Hay, T.A. (2001). Electrochemical cell diagnostics using online impedance measurement, state estimation and data fusion techniques, *Proceedings of 36th Intersociety Energy Conversion Engineering Conference,* Savannah, Georgia.

Goebel, K., Eklund, N. & Bonanni, P. (2006). Fusing Competing Prediction Algorithms for Prognostics, *Proceedings of 2006 IEEE Aerospace Conference*, New York.

Saha, B., Goebel, K., Poll, S. & Christophersen, J. (2009). Prognostics methods for battery health monitoring using a Bayesian framework, *IEEE Transaction on Instrumentation and Measurement*,

vol. 58, no. 2, pp. 291–296.

Perrone, M.P., and Cooper, L.N. (1993). When networks disagree: ensemble methods for hybrid neural networks, *Neural Networks for Speech and Image Processing*. R.J. Mammone, ed., Chapman-Hall, 1993.

Bishop, C.M. (2005). *Neural networks for pattern recognition*. Oxford University Press.

Goel, T., Haftka, R.T., Shyy, W. & Queipo, N.V. (2007). Ensemble of surrogates, *Structural and Multidisciplinary Optimization*, vol. 33, no. 3, pp. 199–216.

Acar, E. & Rais-Rohani, M. (2009). Ensemble of metamodels with optimized weight factors, *Structural and Multidisciplinary Optimization*, vol. 37, no. 3, pp. 279–294.

Hu, J., Yang, Y.D. & Kihara, D. (2006). EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences, *BMC bioinformatics*, vol. 7, no. 342.

Chen, S., Wang, W. & Zuylen, H. (2009). Construct support vector machine ensemble to detect traffic incident, *Expert Systems with Applications*, vol. 36, no. 8, pp. 10976–10986.

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, vol. 53, no. 4, pp. 343–367.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the International Joint Conference on Artificial Intelligence - IJCAI'95'*.

Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine, *Journal of Machine Learning Research*, vol. 1, p211–244.

Smola, A.J. & Schölkopf, B. (2004). A tutorial on support vector regression, *Statistics and Computing*, vol. 14, no. 3, pp. 199–222.

Cernansky, M., Makula, M. & Cernansky, L., (2007). Organization of the state space of a simple recurrent network before and after training on recursive linguistic structures, *Neural Networks*, vol. 20, no. 2, pp. 236–244.

Saxena, A. & Goebel, K. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation, *IEEE, International Conference on Prognostics and Health Management*, Denver, CO, Oct 6-9.

Leibfried, T. (1998). Online monitors keep transformers in service, *IEEE. Computer Applications in Power*, vol. 11, no. 3, pp. 36–42.

Rivera, H.L., Garcia-Souto, J.A. & Sanz, J. (2000) Measurements of mechanical vibrations at magnetic cores of power transformers with fiber-optic interferometric intrinsic sensor, *IEEE Journal on Selected Topics in Quantum Electronics*, vol. 6, no. 5, p788–797.

Xu, H. & Rahman, S. (2005). Decomposition methods for structural reliability analysis, *Probabilistic Engineering Mechanics*, vol. 20, no. 3, pp. 239–250.

Chao Hu: Mr. Hu received his B.E. degree in Engineering Physics from Tsinghua University (Beijing, China) in 2003. He is currently pursuing the Ph.D. degree in mechanical engineering at The University of Maryland, College Park (Maryland, USA). His research interests are system reliability analysis, prognostics and health management (PHM), and battery power and health management of Li-ion battery system.

Byeng D. Youn: Dr. Byeng D. Youn is currently an Assistant Professor in the School of Mechanical and Aerospace Engineering at Seoul National University in South Korea. Dr. Youn is dedicated to well-balanced experimental and simulation studies of system analysis and design and is currently exploring three research avenues: (1) system risk-based design, (2) prognostics and health management (PHM), and (3) energy harvester design. Dr. Youn's research and educational portfolio includes: (i) *six notable awards***,** including the ISSMO/Springer Prize for the Best Young Scientist in 2005 from the International Society of Structural and Multidisciplinary Optimization (ISSMO), (ii) *over one hundred publications* in the area of system risk assessment and design and PHM. His primary applications include Li-ion batteries, consumer electronics, and large-scale engineered systems (e.g., automotives).

Pingfeng Wang: Dr. Wang received his B.E. degree in Mechanical Engineering from The University of Science and Technology (Beijing, China) in 2001, the M.S. degree in Applied Mathematics in Tsinghua University (Beijing, China) in 2006, and the Ph.D. degree in mechanical engineering at the University of Maryland, College Park (Maryland, USA). He is currently an Assistant Professor in the Department of Industrial and Manufacturing Engineering at Wichita State University. His research interests are system reliability analysis, risk-based system design, health monitoring, prognostics and health management (PHM) and structural sensor network design.