Online Anomaly Detection for Hard Disk Drives Based on Mahalanobis Distance

Yu Wang, *Student Member, IEEE*, Qiang Miao, *Senior Member, IEEE*, Eden W. M. Ma, *Member, IEEE*, Kwok-Leung Tsui, and Michael G. Pecht, *Fellow, IEEE*

***Abstract*—A hard disk drive (HDD) failure may cause serious data loss and catastrophic consequences. Online health monitoring provides information about the degradation trend of the HDD, and hence the early warning of failures, which gives us a chance to save the data. This paper developed an approach for HDD anomaly detection using Mahalanobis distance (MD). Critical parameters were selected using failure modes, mechanisms, and effects analysis (FMMEA), and the minimum redundancy maximum relevance (mRMR) method. A self-monitoring, analysis, and reporting technology (SMART) data set is used to evaluate the performance of the developed approach. The result shows that about 67% of the anomalies of failed drives can be detected with zero false alarm rate, and most of them can provide users with at least 20 hours during which to backup the data.**

***Index Terms*—Hard disk drive, Mahalanobis distance, online anomaly detection, self-monitoring, analysis, and reporting technology.**

Acronyms

HDD	hard disk drive
SMART	self-monitoring, analysis, and reporting technology
FDR	failure detection rate
SVM	support vector machine
FAR	false alarm rate
NBEM	naive Bayes expectation-maximization
FSMD	feature selection based Mahalanobis distance
MRMR	minimum redundancy maximum relevance
MD	Mahalanobis distance
M–P	Moore–Penrose
SVD	singular value decomposition
ROC	receiver operating characteristic
FMMEA	failure modes, mechanisms and effects analysis
PoF	physics-of-failure
HDI	head-disk interface
HMM	hidden Markov model
MAD	median absolute deviation

Notation

I	mutual information
$p(x, y)$	joint probability distribution function of X and Y
S	feature space
f_i	ith feature
c	class variable
$\mathbf{X}$	data set
x_{ij}	jth attribute at ith observation
$\mathbf{X}_j$	columns of $\mathbf{X}$
n	the number of attributes
m	the number of observations
$\overline{\mathbf{X}}_j$	mean of $\mathbf{X}_j$
S_j	standard deviation of $\mathbf{X}_j$
MD_i	Mahalanobis distance for ith observation
$\mathbf{C}$	covariance matrix
$\mathbf{C}^{-1}$	inverse of $\mathbf{C}$
$\mathbf{U}$	unitary matrices, left singular vectors of $\mathbf{C}$
$\mathbf{V}$	unitary matrices, right singular vectors of $\mathbf{C}$
$\mathbf{D}$	diagonal matrix, the singular values of $\mathbf{C}$
$diag$	diagonal matrix
$\mathbf{C}^{\dagger}$	Moore-Penrose pseudoinverse matrix of $\mathbf{C}$
ψ	smooth logistic function

Manuscript received November 12, 2011; revised June 30, 2012; accepted July 20, 2012. Date of publication January 30, 2013; date of current version February 27, 2013. This work was fully supported by a Grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU8/CRF/09). Associate Editor: P. Lall.

Y. Wang and E. W. M. Ma are with the Center for Prognostics and System Health Management, City University of Hong Kong, Hong Kong (e-mail: ywang95@student.cityu.edu.hk; eden.wm.ma@cityu.edu.hk).

Q. Miao was with the Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park, MD 20742 USA. He is now with School of Mechanical, Electronic, and Industrial Engineering, University of Electronic Science and Technology of China (e-mail: mqiang@uestc.edu.cn).

K.-L. Tsui is with the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong (e-mail: kltsui@cityu.edu.hk).

M. G. Pecht is with the Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park, MD 20742 USA (e-mail: pecht@calce.umd.edu).

Digital Object Identifier 10.1109/TR.2013.2241204

E	estimator set
T	threshold set
L_n	location estimator
S_n	Scale estimator

I. Introduction

A hard disk drive (HDD) is one of the most common devices for data storage. A survey showed that most information in the world is being stored on HDDs [1]. In some large-scale storage systems, such as enterprise systems, data centers, and Internet service providers, the number of HDDs in a node can easily reach 1000 [2]. The failure of HDDs could not only lead to service downtime, but could also induce serious data loss, and even catastrophic consequences. Therefore, detecting anomalies and providing impending failure warning to users have become a challenge for both HDD manufacturers and users.

The use of health monitoring strategies on HDDs may overcome this problem [3]. Health monitoring is a methodology that assesses the deviation or degradation of a product from its normal condition by fusing sensor data with models [3], [4]. It has the ability to detect the anomalous behaviors of a product online, which is useful to provide warnings of impending failures to the users, as well as information to experts for diagnostics. With an online health monitoring capability in place, HDDs can 1) perform self-correction for logical errors [5] once the anomaly is detected; and 2) remind users to back up their data.

To make online monitoring for HDDs feasible, the monitoring technique should be convenient, inexpensive, and unobtrusive. The health monitoring techniques currently used in HDD can be categorized into three groups. Group A utilizes external sensors, such as accelerometers and acoustic emission sensors, to monitor the evolution of HDD vibration signals from the inception to the end of life under an accelerated life test [6]. However, attaching these sensors outside the HDD does not allow for obtaining degradation signatures, while attaching inside the HDD will violate the requirement to be unobtrusive [6]. Moreover, the sensors are also expensive compared to the cost of the HDD, which increase the monitoring cost. Group B utilizes the log files of storage systems, which collect the error events of the software and hardware deployed in the storage system [2]. The disadvantage of this technique is that it fails to provide the insightful information associated with HDD performance due to lack of close monitoring [7]. Group C is self-monitoring, analysis, and reporting technology (SMART), which is a built-in function for current HDDs. It reports the performance characteristics of HDDs at specified intervals without intrusion. The implementation of SMART is easy because there is no additional hardware requirement, and the parameters can be monitored via the interface between the computer's start-up program (BIOS) and HDD [3], [8]. The key parameters used to predict failures are track-seek retries, read errors, write faults, reallocated sectors, head fly heights, and environmental temperatures [8].

The anomaly detection algorithm for SMART data used by HDD manufacturers is known as the "threshold method". Thresholds of each SMART parameter are set by the manufacturers [8]. If any parameter exceeds the threshold, the SMART system issues a warning that the HDD is likely to fail soon [8]. Because of warranty issues, the manufacturers set the threshold as high as possible to avoid false alarms (i.e., a healthy HDD categorized as a failed HDD) [9]–[11], thereby minimizing the number of field-returned drives [10]. This strategy obscures the actual failure detection rate (FDR). As reported in [11], the FDR of the SMART program was only 3% to 10%.

New algorithms have been developed during the last decade to improve the FDR. Hughes *et al.* [10] found that most SMART parameters were nonparametrically distributed, and applied a non-parametric method, rank-sum test, to replace the threshold method. Hamerly *et al.* [12] developed naive Bayes expectation-maximization (NBEM), a semi-supervised method, to predict drives' failures. The results showed that using three parameters (grown defect count, read soft errors, and seek errors) led to better prediction performance than using all parameters. This result implies that some parameters are not as useful as expected. In [11], several machine learning methods, including multiple-instance naive Bayes, support vector machines, autoclass, and rank-sum test for anomaly detection were compared. Among these methods, support vector machine (SVM) provided the best FDR, 50.6%, when the false alarm rate (FAR) was at 0%. Although SVM obtains the best prediction performance out of all of the above methods, its computation (computational efficiency and memory use) is too expensive for online monitoring. Several approaches including tree augmented naive Bayesian [13], rule-based approach [14], and hidden Markov model (HMM) [15] were proposed to achieve high computational efficiency. The results of [13] showed that they could achieve high computational efficiency, and more than 80% FDR at around 3% FAR. However, at 0% FAR, [13] showed only 20–30% FDR. The rule-based approach [14] cannot bring the FAR below 3% due to the limitation of the model. The HMM-based approach [15] achieves 52% FDR at 0% FAR, which is similar to [11], but fails to get a higher FDR at higher FAR compared to [13] and [14].

In this paper, we develop an online anomaly detection approach, called feature selection based Mahalanobis distance (FSMD), to address both the computational and prediction problems. The implementation of Mahalanobis distance (MD) transforms the multivariate data into univariate data, which enhances the computational efficiency dramatically. The application of a rule-based anomaly detection algorithm to the critical feature subset, which is selected in the data preparation phase, enhances the prediction accuracy.

The remainder of this paper is organized as follows. In Section II, the related theoretical background is introduced. In Section III, the developed approach is presented step by step. To demonstrate the effectiveness of our approach, a real HDD data set is evaluated in Section IV. The conclusions are drawn in the Section V.

II. Theoretical Background

A. Minimum Redundancy Maximum Relevance (mRMR)

The minimum redundancy maximum relevance (mRMR) method uses mutual information, correlations, and distance or similarity scores as feature selection criteria [16]. It is a robust feature selection method that can rank the features based on their relevance to the target, and exclude the redundant features as well. To sum up, this combined method selects the features with maximum relevance and minimum redundancy.

Given two discrete variables X and Y, the mutual information of these two variables can be defined as

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1}$$

where $p(x), p(y)$ are the marginal probability distribution functions of X, and Y respectively.

The relevance between the features and class variables can be achieved by

$$D(S,c)=\frac{1}{|S|}\sum_{f_i\in S}I(f_i;c) \tag{2}$$

The redundancy of all of the features takes the form

$$R(S)=\frac{1}{|S|^2}\sum_{f_i,f_j\in S}I(f_i;f_j) \tag{3}$$

Finally, the mRMR criterion is the combination of these two constraints:

$$Max\,\{D(S,c)-R(S)\} \tag{4}$$

B. Mahalanobis Distance (MD)

Mahalanobis distance (MD) is a generalized distance which is useful for determining the similarity between an unknown sample and a collection of known samples by considering the correlations between the variables [17], [18]. In this study, the data set can be denoted as $\mathbf{X}$, which consists of m observations and n attributes. The columns of $\mathbf{X}$ are attributes (parameters) of the data set that are denoted as $\mathbf{X}_j$, where $j=1,2,\ldots,n$. The value of the ith observation and the jth attribute is denoted by x_{ij}, where $i=1,2,\ldots,m$, and $j=1,2,\ldots,n$. To eliminate the scale effect for different attributes, each individual attribute in every data vector is normalized. The equation is expressed as

$$z_{ij}=\frac{(x_{ij}-\overline{\mathbf{X}}_j)}{S_j},\ i=1,2,\ldots,m,\ j=1,2,\ldots,n, \tag{5}$$

It is noteworthy that the mean and standard deviation are calculated from the healthy data.

$$\bar{X}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij} \tag{6}$$

$$S_j=\sqrt{\frac{\sum_{i=1}^{m}(x_{ij}-\bar{\mathbf{X}}_j)^2}{m-1}} \tag{7}$$

The MD value of each observation is

$$MD_i=\frac{1}{n}\mathbf{z}_i\mathbf{C}^{-1}\mathbf{z}_i^T, \tag{8}$$

where $\mathbf{z}_i=[z_{i1},z_{i2},\ldots,z_{in}]$, $\mathbf{z}_i^T$ is the transpose vector of $\mathbf{z}_i$, and $\mathbf{C}$ is the covariance matrix

$$\mathbf{C}=\frac{1}{(m-1)}\sum_{i=1}^{m}\mathbf{z}_i^T\mathbf{z}_i \tag{9}$$

In practice, the inverse of the matrix may not exist due to the singularity issue, also known as multicollinearity (high correlation among the attributes) [17]. To solve this problem, Han *et al.* [19] proposed a robust method for MD calculation using Moore-Penrose (M-P) pseudoinverse.

C. Moore–Penrose (M–P) Pseudoinverse

M–P pseudoinverse is a generalized inverse calculated by singular value decomposition (SVD) [20].

There exist orthogonal matrices $\mathbf{U}_{p\times p}$ and $\mathbf{V}_{q\times q}$ such that matrix $\mathbf{C}_{p\times q}$ can be decomposed as

$$\mathbf{C}=\mathbf{U}\mathbf{D}\mathbf{V}^T \tag{10}$$

The columns of $\mathbf{U}$ are the left singular vectors, and the columns of $\mathbf{V}$ are the right singular vectors. $\mathbf{D}$ is an $p\times q$ diagonal matrix, which contains all of the singular values of $\mathbf{C}$.

$$\mathbf{D}=\begin{bmatrix}\mathbf{D}_r & 0\\ 0 & 0\end{bmatrix} \tag{11}$$

where $\mathbf{D}_r=diag(\sigma_1,\ldots,\sigma_r)$, $\sigma_1,\sigma_2,\ldots,\sigma_r$ are the non-zero singular values, and $r=rank(\mathbf{C})$.

The pseudo inverse matrix of $\mathbf{C}$ takes the form

$$\mathbf{C}^\dagger=\mathbf{V}\mathbf{D}^\dagger\mathbf{U}^T \tag{12}$$

where "†" means pseudo inverse. $D^\dagger$ takes the form

$$\mathbf{D}^\dagger=\begin{bmatrix}\mathbf{D}_r^{-1} & 0\\ 0 & 0\end{bmatrix} \tag{13}$$

Although the M–P pseudoinverse is mainly used to overcome the multicollinearity problems, it can still be used without them. In this case, the pseudo inverse matrix equals the inverse matrix.

D. Receiver Operating Characteristic (ROC)

The receiver operating characteristic (ROC) curve is the tradeoff plot between failure detection rate (FDR) and false alarm rate (FAR) [21]. The ROC principle can be described by the detection errors (false alarm, and miss detection). A false alarm is known as a false positive or Type I error, which is a statistical error that incorrectly classified the product as unhealthy when it is in fact healthy. Conversely, a miss detection is known as a false negative or Type II error, wherein a product is determined to be healthy when it is not. An example describing

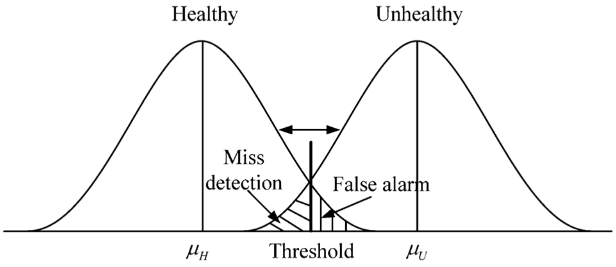

Fig. 1. Classification errors: μ_H is the mean value of healthy group, and μ_U is the mean value of unhealthy group [21].

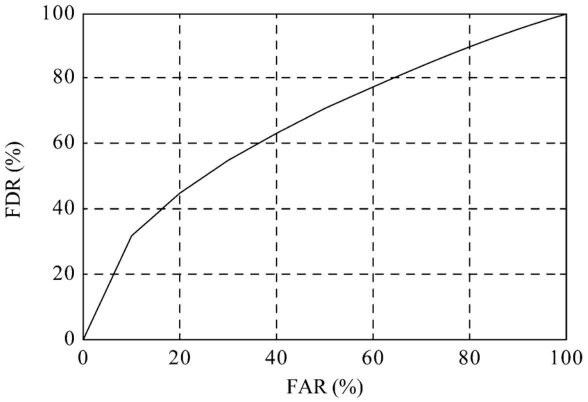

Fig. 2. ROC curve [21].

the relationship between a false alarm and a miss detection is shown in Fig. 1, where the healthy and unhealthy groups are assumed to follow normal distributions with different mean values.

FAR is the ratio of the false alarmed products to the total healthy products.

$$FAR = \frac{n_{fa}}{n_H} \tag{14}$$

where n_{fa} is the number of the false alarmed products, and n_H is the number of total healthy products.

Miss detection rate is the ratio of miss detected products to the total number of unhealthy products. FDR can be obtained by using one minus the miss detection rate, or defined as the ratio of the detected unhealthy products to the total unhealthy products:

$$FDR = 1 - \frac{n_{md}}{n_U}, \text{or} \frac{n_{du}}{n_U} \tag{15}$$

where n_{md} is the number of miss detected products, n_{du} is the number of detected unhealthy products, and n_U is the total number of unhealthy products.

The FAR and FDR can be varied by choosing different thresholds, and the different pairs of FAR and FDR form the ROC curve (Fig. 2).

III. Developed Approach

Fig. 3 shows the scheme of the developed approach, FSMD. SMART data containing both healthy and failed data are collected by the manufacturers. Instead of using all SMART parameters, critical parameters are selected by using a combined feature selection method. In other words, a feature subset is first chosen by failure modes, mechanisms, and effects analysis (FMMEA); then, it will be chosen by the minimum redundancy maximum relevance (mRMR) method. Once the critical parameters have been determined, a baseline Mahalanobis space will be built up using the healthy data of the selected parameters. The failure threshold can be defined by limiting the FAR in the healthy data. Then the same parameters are monitored in the testing system. By comparing with the failure threshold, the anomalies can be detected based on a rule-based method, which will be discussed in Section III-D.

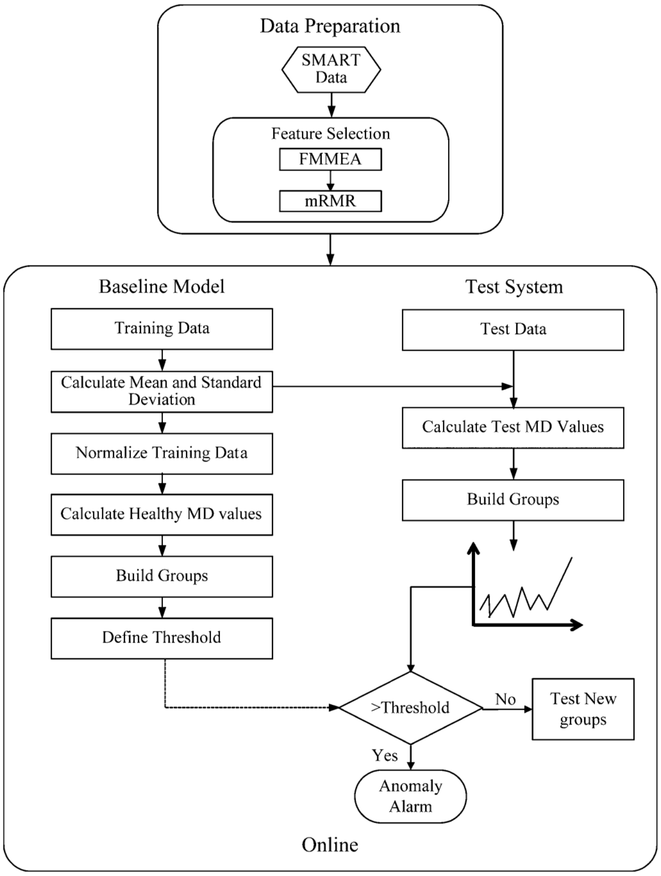

Fig. 3. Scheme of FSMD.

A. Feature Selection

Feature selection is performed during the data preparation stage to determine the critical parameters. A two-step feature selection, including FMMEA and mRMR, is used in this paper.

FMMEA is a physics-of-failure (PoF) based methodology. It identifies the potential failure mechanisms and models for all potential failure modes under the operational and environmental conditions. Then, the failure mechanisms are prioritized according to their severity and likelihood of occurrence so as to determine the failure precursors [3]. The FMMEA investigation on HDD was conducted in [22]. The dominant failure mechanisms were identified as wear out, overstress of head-disk interface (HDI), and resonance of head stack assembly. Those parameters that have a strong correlation with the HDI and head

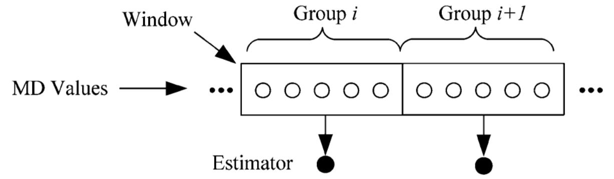

Fig. 4. Window operation for the healthy data.

stack assembly were selected as critical parameters [22]. However, as FMMEA did not take the correlation between the parameters into account, the selected parameters may be highly redundant. Redundant features (parameters) do not increase diagnostic power, but do increase computational cost. Hence, it is necessary to minimize redundancy.

Currently, mRMR is widely used in the field of feature selection and classification. As mRMR ranks all features according to their relevance to the target while excluding the redundancy of features, it is a good solution to reduce the feature space. mRMR is a supervised learning method which needs both the healthy and failed data to build a feature selector. Therefore, the data with FMMEA features that belonged to different classes (healthy and failed) were prepared first. Then, the mutual information between the features and target classes was calculated by (2). Afterwards, the redundancy among the features was calculated by (3). And then, (4) was used to get the scores of all of the features. After ranking the scores, the features at the top of the list were selected as critical features.

B. Baseline Model

Only the reduced features from the previous section were monitored for the healthy drives or testing drives. The baseline Mahalanobis space was built by using the healthy drives' data. Firstly, the parameters of the healthy data were normalized to eliminate the scale effect using (5). The covariance matrix, **C**, was calculated by (9) using the normalized values. To eliminate the multicollinearity problem caused by the strong correlation, the inverse matrix of **C** was determined by the M–P pseudoinverse. SVD (10)–(13) was used to calculate the M-P pseudoinverse. Then, the MD values for healthy drives were calculated using (8). In practice, the transient peaks in time series would induce false alarm, and should be avoided. To eliminate the transient peak, a fixed-size window was used to group the MD values in each drive, and the estimators were used to represent the group of data in each window. The process to build the groups and get the estimators in a healthy HDD is shown in Fig. 4. These estimators in the training HDDs constitute the baseline groups.

In the industry, minimizing FAR is preferable for manufacturers due to warranty issues [9]–[11]. The threshold for the baseline model is set to 0% FAR. The FAR is the ratio of the number of false alarmed HDDs to the total number of the healthy HDDs that participate in building the baseline model.

C. Estimators

HDD anomalies can be identified as either location changes (Fig. 5(a)), or scale changes (Fig. 5(b)). To capture these anomalies, this paper introduces four robust estimators: median, median absolute deviation (MAD), location M-estimator, and scale M-estimator. The former two estimators are commonly used in the literature, and the latter two estimators are proposed by P. J. Rousseeuw *et al.* [23] to give the robust estimates in small samples ($n \geq 4$).

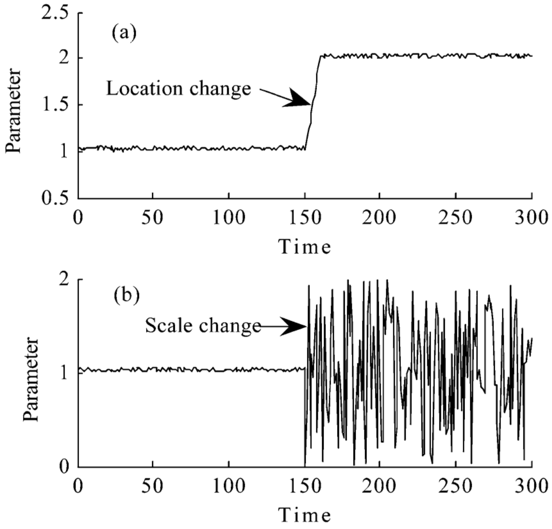

Fig. 5. Two types of anomalies. (a) Location change; (b) Scale change.

Compared with the mean and mode, median is a more robust measure of location [24]. Given there are n observations, $X = (x_1, x_2, \ldots, x_n)$, in a window, the median value is defined by

$$\mathrm{median}(X) = \begin{cases} x_{\frac{n+1}{2}} & \text{when } n \text{ is odd,} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{when } n \text{ is even} \end{cases} \tag{16}$$

MAD is a robust measure of scale (variability) in a data series. It is defined by

$$\mathrm{MAD}(X) = 1.4826 \mathrm{median} \left| x_i - \mathrm{median}(X) \right| \tag{17}$$

The M-estimators are more advanced estimators for both location and scale, which use the median and MAD in their constructions [23]. P. J. Rousseeuw *et al.* [23] proposed a location M-estimator using a smooth logistic ψ function

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - L_n}{S_n}\right) = 0 \tag{18}$$

where S_n takes $\mathrm{MAD}(X)$, and L_n is the location estimator, which can be calculated by Newton-Raphson algorithm with starting initial location estimate $L_n^{(0)} = \mathrm{median}(X)$. $L_n^{(k)}$ can be updated by its previous step $L_n^{(k-1)}$ until it converges.

$$L_n^{(k)} = L_n^{(k-1)} + S_n \frac{\sum_{i=1}^{n} \psi\left(\frac{\left(x_i - L_n^{(k-1)}\right)}{S_n}\right)}{\sum_{i=1}^{n} \psi'\left(\frac{\left(x_i - L_n^{(k-1)}\right)}{S_n}\right)} \tag{19}$$

The logistic ψ function takes the form

$$\psi(x) = \frac{e^x - 1}{e^x + 1} \tag{20}$$

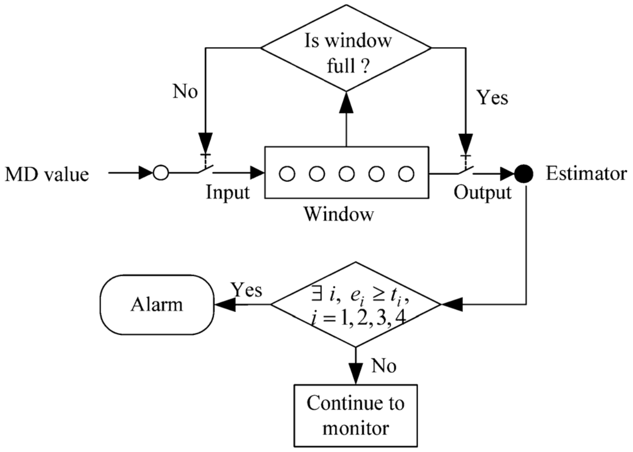

Fig. 6. Window operation during test process.

For M-estimators of scale, an auxiliary location estimate $L_n(x_1, x_2, \ldots, x_n)$ is used, and takes $L_n(X) = \mathrm{median}(X)$ [23].

$$\sum_{i=1}^{n} \rho\left(\frac{x_i - L_n}{S_n}\right) = \beta \tag{21}$$

$\rho(x)$ takes the form

$$\rho(x) = \psi^2\left(\frac{x}{0.3739}\right) \tag{22}$$

To compute S_n, the iterative steps are given with the starting initial scale estimate $S_n^{(0)} = \mathrm{MAD}(X)$.

$$S_n^{(k)} = S_n^{(k-1)} \sqrt{2 \sum_{i=1}^{n} \rho\left(\frac{(x_i - L_n)}{S_n^{(k-1)}}\right)} \tag{23}$$

D. Rule-Based Anomaly Detection

Fig. 6 shows an operation in the testing process. The parameters' mean, standard deviation, and the covariance matrix of healthy drives are used to calculate the MD value of each observation in the testing set.

Then, the consecutive MD values are fed into the fixed-size window. When the window is full, these MD values are used to build the four estimators. Instead of using the individual estimator to determine the drives' status, a logical "OR" rule based on the four estimators is used to make the decision.

The estimator set in a window k is denoted as $E^k = (e_1^k, e_2^k, e_3^k, e_4^k)$, where e_1, e_2, e_3, and e_4 stand for median, MAD, location M-estimator, and scale M-estimator, respectively. The threshold set for these four estimators is denoted as $T = (t_1, t_2, t_3, t_4)$.

The "OR" rule can be explained as: *if any e_i^k exceeds t_i* (24), *the anomaly happens, and the status of this HDD is determined as failed* $(C = 1)$; *otherwise, the HDD is determined as healthy* $(C = 0)$.

$$\exists i,\ e_i^k \geq t_i,\ i = 1, 2, 3, 4 \tag{24}$$

Based on the previous discussion, the detailed algorithm can be described as follows.

Algorithm "OR" rule-based anomaly detection

```
Algorithm
BEGIN
1: Input: The new observations, MD values, and four estimators in a fixed-size window, as well as failure thresholds.
2: Output: The decision of the drive's status $C$, and the failure time $F_T$.
3: Step 1: Calculate MD values, and estimators $e_i$, $i = 1, 2, 3, 4$.
4: Calculate the MD values.
5: Wait until the window is full.
6: Calculate the four estimators $e_i$, $i = 1, 2, 3, 4$.
7: Step 2: Make the decision for the drive's status.
8: Use the four estimators $(e_i, i = 1, 2, 3, 4)$ to compare with their thresholds $T = (t_1, t_2, t_3, t_4)$.
9: if $\exists i, e_i \geq t_i, i = 1, 2, 3, 4$ then execute 10 and 11.
10: The drive is determined as failed, $C = 1$.
11: An alarm will be reported, and failure time is detected, $F_T$.
12: Else execute 13 and 14.
13: The drive is determined as healthy, $C = 0$.
14: The monitoring process continues until the drive's anomaly being detected.
15: End if.
16: Return $(C, F_T)$.
17: END
```

E. Performance Evaluation

To evaluate the prediction performance provided by our developed approach, three metrics are introduced: prediction accuracy, ROC curve, and time before failure. Brief descriptions of implemented metrics are as follows.

- Prediction accuracy—This metric refers to the ratio of the number of detected drives, including both healthy and failed, to total drives.
- ROC curve—This metric is used to show the tradeoff between FDR and FAR. To draw a ROC curve, the thresholds of four estimators are adjusted to obtain different pairs of FDR and FAR.
- Time before failure—This metric is defined by the time interval between reporting the alarm and the drive can not work at all (complete failure) [11]. A reasonable value of this metric can guarantee the data being backed up in time. The goal of SMART designed by HDD manufacturers is to provide 24 hours warning-time before drive failure [10].

IV. Results and Discussions

A. Brief Description of Smart Data Set

The developed approach is validated by using a SMART data set, which includes 369 drives from one model, in which 178 drives were labeled as good (healthy) and 191 drives were labeled as failed [25]. The healthy drives were from a reliability test [11] by the manufacturer. The failed drives were field returned. All drives were verified by the manufacturer at the end of the test; the healthy drives could run well and the failed drives could not. The most recent 300 samples (observations) were recorded in each drive, and the time interval between two samples was 2 hours. Only the last 600 hours of data were recorded; when the time exceeded 600 hours, the data were then overwritten. Some failed drives had less than 300 samples because they were not able to survive 600 hours of operation. Each sample contained 60 performance characteristic attributes, and some additional attributes such as the drive's serial number and power-on-hours [11].

B. Selected Features

A preliminary examination was executed before using the data. Attributes whose measurements were all zeros were abandoned.

As discussed previously, FMMEA and mRMR were used to determine the critical parameters. FMMEA determined 47 parameters including head flying heights, read/write errors, primary defect, growth defect, sectors read/write, and servo errors as critical features. Detailed information can be found in [22].

In mRMR, the last 50 samples from 1/3 of the failed drives were used as the abnormal group based on the assumption that the last 50 samples could demonstrate the abnormal signature. All samples from 1/3 of the healthy drives were used as the normal group. This configuration is similar to the configuration in [11], in which 1/3 of all drives (both healthy and unhealthy) were used in feature selection. Then these data were fed into the mRMR software developed by Peng [26] to calculate the mutual information and get the scores of each feature using (4). The mRMR result for the FMMEA parameters is shown in Fig. 7. The normalized scores were used to evaluate the contributions of the features. The bigger scores represent the bigger contribution of these features. It is found that, starting from the 31st feature, the score of FMMEA parameters stopped increasing. This result indicates that the remaining features have no contribution at all. Thus, these 31 features were selected as critical parameters, denoted as FMMEA-mRMR parameters. All of the following analysis and discussions were based on the FMMEA-mRMR parameters.

C. FSMD-Based Anomaly Detection

Randomly select 60% (107/178) of the healthy drives using FMMEA-mRMR parameters to build the baseline Mahalanobis space. The window size was set at 5 by referring to [11]. This window size also satisfies the requirement for getting the estimators. The thresholds for the four estimators were defined by searching this baseline space with constraining FAR at 0%. The remaining 40% of the healthy drives and all of the failed drives were used as a testing set. Ten trials are prepared. At each trial, 60% of the healthy drives were randomly assigned to the training set, while the remaining drives were assigned to the testing set. To verify whether the "OR" rule-based method is superior to purely using the location estimators (e_1 and e_3) or scale estimators (e_2 and e_4), the comparison of these three sets were also conducted. All results below are averaged over 10 trials.

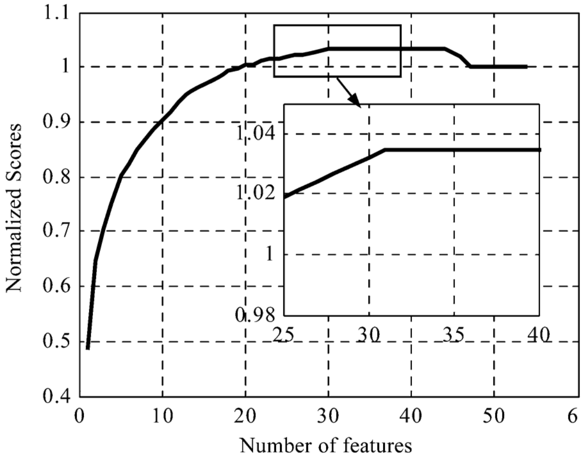

Fig. 7. mRMR result for FMMEA parameters.

TABLE I
Using the Trained Thresholds to Detect Anomalies

Estimator set	Prediction accuracy* (%)
'OR' rule-based	79.2
Location estimator	72.8
Scale estimator	65.2

* Only the tested drives participate in calculating the prediction accuracy.

Table I shows the prediction accuracy using the three estimator sets. It can be observed that using the "OR" rule-based method achieved the prediction accuracy of 79.2%, which outperforms purely using the location estimator set or the scale estimator set.

To get a comprehensive comparison, the thresholds were adjusted to obtain the different pairs of FAR and FDR to form the ROC curve. The results were illustrated in Fig. 8. It is found that using the "OR" rule-based method can achieve 67% FDR at zero FAR, which is 15% better than purely using location estimator, and 14% better than purely using scale estimator. This result indicates that both types of anomalies were encountered in the failed drives. It is also proved that using the "OR" rule-based method is necessary to achieve the high prediction accuracy.

To explore the prediction capability in terms of time before failure, all the failed drives, at the point 67% FDR, zero FAR, were inspected. Fig. 9 illustrates how to calculate the time before failure in an individual drive (drive #2).

The first anomaly of this drive was detected by e_1, e_2, and e_3 simultaneously. e_1 was used to illustrate the calculation. As shown in Fig. 9, the first anomaly was detected at the 13th group (corresponding to the 65th observation) when the testing system reported an alarm. From the first anomaly being detected in the drive to the end of its life, drive #2 spanned 30 groups, corresponding to 300 hours. In other words, the user

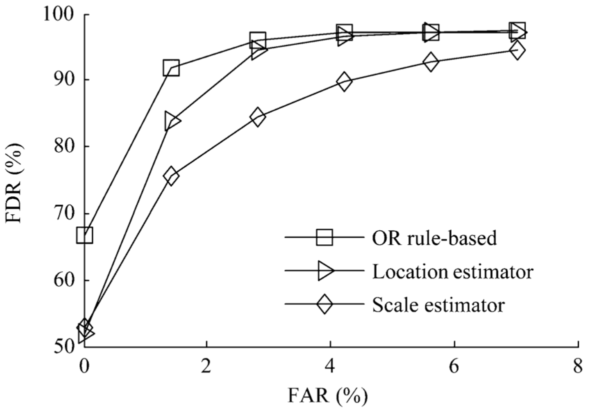

Fig. 8. Performance of "OR" rule-based method, location estimator, and scale estimator on HDD failure prediction.

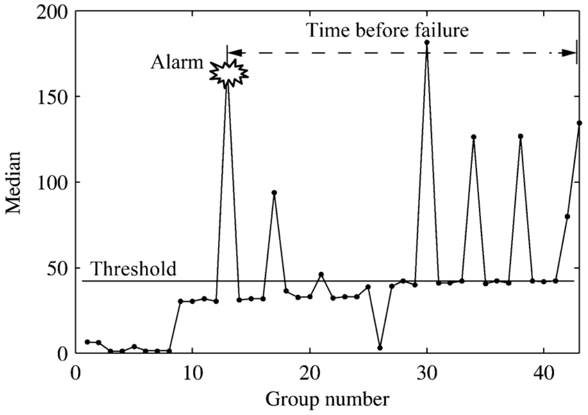

Fig. 9. Anomaly detection for drive #2.

TABLE II
SUMMARY OF TIME BEFORE FAILURES FOR FAILED DRIVES

Time before failure	No. of HDDs	Percentage (%)
Non-detectable	63	32.98
≧ 0 hours	128	67.02
≧ 10 hours	117	61.25
≧ 20 hours	107	56.02
≧ 30 hours	97	50.78
≧ 40 hours	78	40.83

has a 300-hour margin to back up data when the system reports the alarm. Table II summarized the time before failures for all the failed drives. As seen from this table, 56% (107/191) of the failed drives could be detected before the last 20 hours of their lives, which is sufficient to back up data.

D. Comparison with the Results of [11]

As previously discussed, [9]–[12] worked to improve the FDR. SVM using the radial kernel function showed the best performance (50.6% FDR at 0% FAR) in anomaly detection among these studies. The SVM-based approach used in [11] follows three steps: 1) 25 features were selected by a combination of the z-score and reverse arrangement tests; 2) ten trials for SVM classification were conducted, and the HDDs were randomly assigned into training and testing sets for each trial; and 3) an averaged ROC curve of the ten trials was performed to evaluate the classification results.

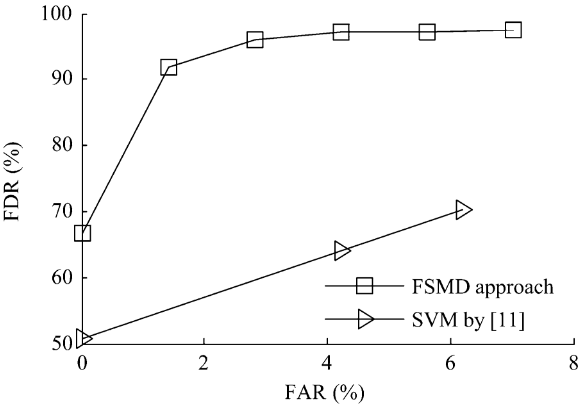

Fig. 10. ROC comparison between FSMD and [11].

TABLE III
COMPARISON BETWEEN FSMD AND SVM

Approaches	Time consumption (Mins.)
FSMD	4.3
SVM by [11]	17983 (3560*)

* The result was reproduced using the same platform with FSMD.

The comparison of the FSMD approach and the approach in [11] is shown in Fig. 10. As seen in this figure, the overall performance of the FSMD approach is much better than [11]. In the lower FAR region, manufacturers are more concerned with showing that the FSMD approach also outperformed the SVM proposed by [11]. Especially at zero FAR, FSMD shows about 17% higher than SVM.

Table III compares the computational efficiency of the FSMD approach and the SVM proposed by [11]. The FSMD approach showed high computational efficiency, taking only 4.3 minutes (using the Matlab 7.1 environment on an Intel Dual Core2 Processor running at 3.00 GHz, and 3.25 GB RAM.), including the training and testing processes for all the HDDs. In contrast, the SVM-based approach used in [11] required 17,983 minutes, around 12 days, for the computation. Due to the different platforms used by FSMD and the SVM in [11], the work in [11] was reproduced. The result shows the time consumption was 3,560 minutes, which is about 2.5 days, 800 times to FSMD.

This paper discussed the training and testing as a whole. The reason is that each new HDD system may have new characteristics, and thus new observations corresponding to these new characteristics should be added into the training set. Although this problem was not addressed in our paper, our approach should be an easier way to implement this strategy than SVM due to its simplicity, and its high computational efficiency.

In terms of time before failure, FSMD can detect about 56% (107/191) failed drive with more than 20 hours of margin, which is much better than [11]. Concerning the time before failure of more than 40 hours, our result is a little worse than [11]. However, 20 hours of margin should be sufficient to back up data, and is also very close to the SMRAT's requirement (24 hours). Actually, there is a tradeoff between the time before failure and prediction accuracy. In the future work, this issue will be considered as a key factor in the prediction model.

V. Conclusions

We developed an online anomaly detection approach called feature selection based Mahalanobis distance (FSMD) for HDD. To begin with, FMMEA was used to select a set of features related to potential failure mechanisms. A mRMR approach was then applied to reduce the redundant features among the FMMEA parameters. The critical parameters selected by mRMR were subsequently used by the MD-based monitoring system for on-line anomaly detection. A SMART data set was used to validate this approach.

As discussed earlier, the SVM-based approach, though computationally expensive, achieved a good prediction performance. Therefore, the developed approach was compared with it. The comparison shows that FSMD can achieve higher prediction performance, and requires less computation time. FSMD was able to provide a notification of 20 hours in advance for 56% (107/191) of the failed drives. Moreover, at 0% FAR, FSMD achieved 67% FDR, which is about 17% higher than that of the SVM-based approach, and about 60% higher than the “threshold method” used by HDD manufacturers. In terms of computational efficiency, FSMD takes only 4.3 minutes on a personal computer, nearly 1/800th the time of the SVM-based approach proposed by [11].

To conclude, the FSMD approach developed in this paper has two advantages. First, it enhances the FDR and FAR significantly, which benefits users by providing the warnings of impending HDD failures, allowing users to back up their data in time; and it benefits manufacturers by reducing warranty issues, and hence lowering the costs. Second, FSMD is computationally efficient, which makes it a useful way of implementing online anomaly detection for industrial applications.

References

[1] M. Hilbert and P. Lopez, “The world’s technological capacity to store, communicate, and compute information,” *Science*, vol. 332, no. 60, pp. 60–65, April 2011.

[2] W. Jiang, C. Hu, Y. Zhou, and A. Kanevsky, “Are disks the dominant contributor for storage failures? A comprehensive study of storage subsystem failure characteristics,” *ACM Trans. Storage*, vol. 4, no. 3, pp. 7:1–7:25, October 2008.

[3] M. Pecht, *Prognostics and Health Management of Electronics*. New York, NY, USA: Wiley-Interscience, 2008.

[4] N. Vichare and M. Pecht, “Prognostics and health management of electronics,” *IEEE Trans. Components Packag. Technol.*, vol. 29, no. 1, pp. 222–229, Mar. 2006.

[5] S. X. Wang and A. M. Taratorin, *Magnetic Information Storage Technology*. San Diego, CA, USA: Academic Press, 1999.

[6] S. Kamarthi, A. Zeid, and Y. Bagul, “Assessment of current health of hard disk drives,” in *Proc. IEEE Int. Conf. Autom. Sci. Eng.*, Bangalore, India, Aug. 2009.

[7] E. Pinheiro, W. Weber, and L. A. Barroso, “Failure trends in a large disk drive population,” in *Proc. 5th USENIX Conf. File Storage Technol. (FAST)*, Lose Angeles, CA, USA, Feb. 2007.

[8] R. K. Henry, “Monitoring PC Hardware Sounds in Linux Systems Using the Daubechies D4 Wavelet,” M.A. thesis, Dept. Comput. Sci., East Tennessee State Univ., Nashville, TN, USA, 2005.

[9] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, “Hard drive failure prediction using non-parametric statistical methods,” in *Proc. Int. Conf. Artif. Neural Netw. ICANN* , Istanbul, Turkey, Jun. 2003.

[10] G. F. Hughes, J. F. Murray, K. Kreutz-Delgado, and C. Elkan, “Improved disk-drive failure warnings,” *IEEE Trans. Rel.*, vol. 51, no. 3, pp. 350–357, Sep. 2002.

[11] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, “Machine learning methods for predicting failures in hard drives: A multiple instance application,” *J. Mach. Learn. Res.*, vol. 6, pp. 783–816, 2005.

[12] G. Hamerly and C. Elkan, “Bayesian approaches to failure prediction for disk drives,” in *Proc. Eighteenth Int. Conf. Mach. Learn. (ICML’01)*, Jun. 2001.

[13] Y. Tan and X. Gu, “On predictability of system anomalies in real world,” in *Proc. 18th Annu. IEEE/ACM Int. Symp. Modeling, Anal., Simulation Comput. Telecommun. Syst.*, Miami Beach, FL, USA, Aug. 2010.

[14] V. Agrawal, C. Bhattacharyya, T. Niranjan, and S. Susarla, “Discovering rules from disk events for predicting hard drive failures,” in *Proc. Int. Conf. Mach. Learn. Appl.*, Miami Beach, Florida, Dec. 2009.

[15] Y. Zhao, X. Liu, S. Gan, and W. Zheng, “Predicting disk failure with HMM- and HSMM-based approaches,” in *Proc. 10th Ind. Conf. Adv. Data Mining. Appl., Theoretical Aspects (ICDM 2010)*, Berlin, Germany, 2010.

[16] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal., Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[17] G. Taguchi and R. Jugulum, *The Mahalanobis-Taguchi Strategy: A Pattern Technology System*. New York, NY, USA: Wiley Press, May 2002.

[18] G. Niu, S. Singh, S. W. Hollandc, and M. Pecht, “Health monitoring of electronic products based on Mahalanobis distance and Weibull decision metrics,” *Microelectron. Rel.*, vol. 51, pp. 279–284, Feb. 2011.

[19] Y. Han, “A Study on Robust Optimization and Diagnostic Analysis of Multidimensional System Based on MTS,” Ph.D. dissertation, Dept. Management, Tian Jin Univ., Tian Jin, China, 2007.

[20] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*. New York, NY, USA: Springer-Verlag, 2003.

[21] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, pp. 861–874, 2006.

[22] Y. Wang, Q. Miao, and M. Pecht, “Health monitoring of hard disk drive based on Mahalanobis distance,” in *Proc. IEEE 2011 Prognostics Health Manag. Conf. (PHM-2011)*, Shenzhen, China, May 2011, pp. 23–25.

[23] P. J. Rousseeuw and S. Verboven, “Robust estimation in very small samples,” *Comput. Statist. Data Anal.*, vol. 40, pp. 741–758, 2002.

[24] R. A. Maronna, R. D. Martin, and V. J. Yohai, *Robust Statistics: Theory and Methods*. Chichester, U.K.: John Wiley & Sons, Jun. 2006.

[25] G. F. Hughes, S.M.A.R.T. Data Set. [Online]. Available: http://cmrr.ucsd.edu/people/hughes/smart/dataset/

[26] H. Peng, mRMR (minimum Redundancy Maximum Relevance Feature Selection) [Online]. Available: http://penglab.janelia.org/proj/mRMR/

Yu Wang (S’12) received the B.S. degree in mechanical design and manufacturing automation from Xi’an University of Technology, Xi’an, China, and the M.S. degree in manufacturing engineering and automation from Xi’an Jiao Tong University, Xi’an, China. He is currently working toward the Ph.D. degree in Center for Prognostics and System Health Management at the City University of Hong Kong, Hong Kong.

Qiang Miao (M’06–SM’12) received B.E. and M.S. degrees from Beijing University of Aeronautics and Astronautics, Beijing, China, and the Ph.D. from the University of Toronto, Toronto, ON, Canada, in 2005.

He was working as a Research Associate Professor at CALCE, Department of Mechanical Engineering, University of Maryland, College Park, USA. He is currently a Professor of the School of Mechanical, Electronic, and Industrial Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China. His current research interests include machinery condition monitoring, reliability engineering, and maintenance decision-making. In 2011, he was selected as the Reserve Candidate of Sichuan Province Academic and Technology Leadership, and the New Century Excellent Talents in University. He has published more than 40 research papers.

Eden W. M. Ma (M’08) received the B.Eng. and Ph.D. degrees in electronics engineering from City University of Hong Kong, Hong Kong.

Her research interests are in data mining, clustering, feature selection and extraction, quality control, reliability, and prognostics and system health management.

Kwok-Leung Tsui received the B.Sc. and M.Sc. degrees from the Chinese University of Hong Kong, Hong Kong. He received the Ph.D. degree from the University of Wisconsin, Madison, USA.

He is well known and highly respected in the Applied Statistics community. He is one of the leading experts in the area of experimental design, and a pioneer in Robust Designs and Taguchi methods.

Dr. Tsui was awarded the National Science Foundation Young Investigator Award, and was elected Fellow of the American Statistical Association.

Michael G. Pecht (S'78–M'83–SM'90–F'92) received the M.S. degree in electrical engineering, and the M.S. and Ph.D. degrees in engineering mechanics from the University of Wisconsin-Madison, Madison, USA.

He is currently a Visiting Professor with City University of Hong Kong, Kowloon, Hong Kong. He is also a Chair Professor in mechanical engineering and a Professor in applied mathematics with the University of Maryland, College Park, USA. He is the founder of the Center for Advanced Life Cycle Engineering, University of Maryland, which is funded by over 150 of the world's leading electronics companies at more than six million U.S. dollars per year. He has written more than 20 books, and over 400 technical articles. He consults for 22 major international companies, providing expertise in strategic planning, design, test, intellectual property, and risk assessment of electronic products.

Dr. Pecht is a Professional Engineer, an ASME Fellow, a SAE Fellow, and an IMAPS Fellow. He served as Chief Editor of the IEEE TRANSACTIONS ON RELIABILITY for eight years, and on the advisory board of IEEE Spectrum. He is Chief Editor for *Microelectronics Reliability*, and an Associate Editor for the IEEE TRANSACTIONS ON COMPONENTS AND PACKAGING TECHNOLOGY. He was the recipient of the IEEE Exceptional Technical Achievement Award in 2010; the IEEE Reliability Society's Lifetime Achievement Award in 2008, the highest reliability honor; the European Micro and Nano-Reliability Award for outstanding contributions to reliability research; the 3M Research Award for electronics packaging; and the IMAPS William D. Ashman Memorial Achievement Award for his contributions in electronic reliability analysis.