Received 19 September 2022, accepted 16 October 2022, date of publication 25 October 2022, date of current version 8 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3216885

RESEARCH ARTICLE

A Genetic Algorithm and PCA-Based Feature Selection to Improve the Failure Diagnosis Performance of Railway Vehicle Doors

GIL HYUN KANG[1], HO CHEOL KI[1], SANG HYUN AN[2], JUHEE CHOI[3], (Member, IEEE), AND CHUL SU KIM[4]

[1] Industry-Academic Cooperation Foundation, Korea National University of Transportation, Uiwang 16106, South Korea
[2] Department of Industrial Engineering, Ajou University, Suwon 16499, South Korea
[3] Department of Smart Information and Communication Engineering, Sangmyung University, Cheonan 31066, South Korea
[4] School of Railroad Engineering, Korea National University of Transportation, Uiwang 16106, South Korea

Corresponding author: Chul Su Kim (chalskim@ut.ac.kr)

This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) by the Ministry of Land, Infrastructure and Transport, under Grant 22QPWO-C152223-04.

ABSTRACT The failure diagnosis of railway vehicle door system is carried out using a test bench and machine learning software for the fast and accurate classification. The signal length deviation exists in actual collected data of normal operation and abnormal failures with a time delay. The traditional data multi-segmentation technique for feature extraction has shortcomings by assuming that the measured time-based signals have the same operating time. However, the uniform data segmentation has a difficulty due to the deviation of measured data length. A method of converting time-based data into position-based data was performed to overcome the deviation problem. A method of optimized single-zone data using a genetic algorithm was proposed to improve the classification performance and to reduce computation time, instead of existing the multi-segmentation technique. A principal component analysis-based feature dimensional reduction with explained variance ratio was used to reduce the effect from multi-collinearity of features. Finally, the combination of the proposed methods was compared with individual methods to validate the classification performance by using support vector machine and other classifiers. It was confirmed that the proposed combination method shows the highest classification accuracy of 99.84%.

INDEX TERMS Failure diagnosis, genetic algorithm, principal component analysis, railway vehicle doors.

I. INTRODUCTION

Railway vehicle door system (RVDS) is a device that requires a high reliability because the train delay by door failures causes a severe congestion in train operation. RVDS is classified into an air type and an electric type by its driving power source. It can also be divided into a pocket sliding, an outside sliding, and a plug sliding door depending on the movement of door panels. It is necessary to use sensors appropriate for the machine learning (ML)-based failure diagnosis by door type [1]. Air type doors are driven by air pressure cylinders, so air flow sensors, displacement sensors, and pressure sensors are widely used [2], [3], [4]. In the case of plug doors, the sound sensor and pressure sensor are used to diagnose the failure [5], [6]. On the other hand, electric type doors use dc motor and screw device to convert the rotational force of into linear motion to move the doors back and forth. Consequently, current sensors, voltage sensors, encoder, and other sensors are used to collect the signals of the doors. The time-based signal data are collected from these sensors, modeled after feature design, and faults are classified for failure diagnosis [7], [8].

Before use of ML-based methods, researches related to fault diagnosis of door system have been performed in reliability centered maintenance using fault tree [9], Bayesian network [10], [11], fuzzy classification [12], and reliability

The associate editor coordinating the review of this manuscript and approving it for publication was Zhigang Liu.

assessment method [13], [14], [15]. The study of failure diagnosis typically classified as model-based approach and data-based approach. Model-based approaches of electric door began in the 1990s with the study of using parameter estimation of motor control [16]. An alternative method, the Bond graph was used for modeling to detect and isolate failures of the system [17], [18]. Data-based approach is a data-driven modelling method for which is based on artificial intelligence (AI) and ML. Recent advances in communication technology, large amounts of data in onboard computer of railway vehicle can be transmitted to ground computers wirelessly, increasing the number of research related to ML and AI of door failure classification [19], [20]. In addition to this, the time series data from key devices requiring reliability can also be used to trace deterioration by using regression techniques to predict a remaining useful life of parts [21], [22].

Many studies have chosen the current signal of the motor from the parameters to achieve acceptable performance. Then, the measured data were segmented into characterized divisions for feature extraction based on the assumption that the measured data lengths are uniform. This technique generally segments the data into 3 to 5 zones, e.g. acceleration, constant speed, and deceleration, which are the physical criteria of operation commands. Recent work has obtained high classification performance of more than 95% as a method for feature selection by linear discriminant analysis (LDA) using 13 statistical features from 3 zone segmentation of motor current data [7]. In contrast, a plug door research case, a multi-segmentation of 5 divisions with current, velocity, and position of 3 variables using a random forest classifier, however a relatively low accuracy of less than 85% was achieved [8]. Therefore, it is needed to improve the classification performance of failure diagnosis of RVDS especially with fast computation and high accuracy in feature extraction and feature selection phase.

The methodology for feature selection could be used to enable better classification of predictors for feature dimensional reduction [23], [24], [25], [26]. Comparative study of the classification performance of ML selected features by Fischer's discriminant ratio (FDR) and Pearson correlation coefficient (PCC) with AI method of Convolution Neural Network (CNN) shows that ML is superior to CNN of a low prediction accuracy of around 80% [7]. Aforementioned, multi-segmentation technique can cause multi-collinearity problems of extracted features. In addition, the uniform data segmentation is not easy in reality because most of the time delay occurs in the event of a failure. To access this problem, an alignment of unequal length data is generally required. Typical data alignment techniques are Euclidean, dynamic time warping, uniform scaling, and scaled and warped matching [27], [28]. These techniques are applicable when there is a small-time deviation like normal operation condition. However, if door operation interrupted by unexpected events of several seconds time delay, such as door reopen failure case, it is difficult to divide into uniform data zones because of a big data length deviation. In this case, statistical errors can occur even the entire sequence is rearranged by using those techniques. It is difficult to use differential gradient values for segmenting and some long data needs to be discarded. Thus, the uniform data segmentation increases the amount of data thrown away. Therefore, the remaining problems can be summarized as follows:

First, there is a practical difficulty of separating data by uniform zones, because of the time-delayed abnormal data lengths and different characteristics of failure data; Secondly, the multi-collinearity problem, which exists the in high-dimensional features, and the problem of longer computation time due to repeating feature extraction. To overcome these problems, new methods, which have high failure classification performance and can reduce computation time, were developed in this work. The proposed methods could be summarized as follows:

First, to reduce feature dimension without discarding measurement data, a preprocessing technique of data conversion was developed. That could convert the collected time-based data into position-based data, which based on the door stroke distance of 650mm. Secondly, to reduce the repeating feature extraction calculation of multi-segmentation method, the estimation of optimized single data zone (OSDZ) using a Genetic Algorithm (GA) was introduced. The data conversion technique, the GA, and the principal component analysis (PCA)-based feature dimensional reduction [29], [30] are presented to improve the model performance in this study. The combination of proposed techniques was compared with an existing feature extraction and selection method with FDR and PCC [7], [31]. Several ML classification models were used to compare the performance. Those included SVM [32], [33], [34], [35], k-Nearest Neighbors (kNN) [36], [37] and others, which are typical classification models commonly used in ML for failure diagnosis. In this work, MATLAB of high-level language with the latest toolbox was used for programming.

The conclusions obtained from comparing the failure classification performance of the proposed and existing methods are as follows:

1) The use of converted position-based data shows enough high classification accuracy with current values only instead of using current, voltage and speed of high dimensional data in terms of variable selection.
2) The use of an OSDZ using a GA instead of the 3 data zone segmentation technique for feature extraction improves the failure diagnosis accuracy, solves the multi-collinearity problem, and reduces the computation time.
3) The proposed PCA-based feature dimensional reduction algorithm was found to be more efficient and accurate than existing LDA by using FDR with PCC.

As a result, the preprocessing technique for position-based data transformation, feature extracting from the optimal single section using a GA and feature selection with PCA algorithm could be applied to fast failure diagnosis of RVDS.

II. DOOR TEST AND MACHINE LEARNING-BASED FAILURE DIAGNOSIS

A. RAILWAY VEHICLE DOOR TEST AND CHARACTERISTICS OF FAILURE DATA

The device used for the test is used for the door system development of commuters in Seoul metropolitan area in Korea, as a widely used pocket sliding type. The door unit consists of a dc motor, a door control unit(DCU), and rollers and bearings to help support and move doors, transfer spindles and nuts, and door panels etc. For the door operation of an electric vehicle, a signal is applied to the DCU of the vehicle through a train communication cable when the driver operates the door open switch in the driver's cabin. In the DCU receiving the open signal, the DC 100V power supply motor is rotated through its own control device, and the drive screw is rotated by the motor shaft and the coupler. Eventually, the door panel connected to the screw nut is opened. Then, the passenger moves. On the contrary, when the door closing switch is pressed, the direction of rotation of the motor is changed by the DCU, and the door panel is moved in reverse. This test bench is a device similar to the Bombardier and JR East [38], [39], [40], [41]. The door test bench and the operating parts used to obtain failure data are shown in Figure 1. The train door model is shown in the lower pictures in Figure 1. The lower left figure shows different speed changes in the open and close operation of the door panel. At the beginning of the open operation, the motor is quickly rotated to accelerate so that the panel opens quickly, and then the open operation continues at a constant speed for a certain distance. Finally, the door panel is decelerated to enter the pocket and stop. In order to close the door again after stopping for a predetermined time, the door is closed again by acceleration, constant speed, and deceleration control. And the lower right figure shows typically using DC motor with permanent magnets. The model is sufficiently accurate, and the stator is a permanent magnet, so the voltage (V), current (A) of the amateur coil and motor speed (ω) can be used as parameters [16]. In this study, instead of diagnosing faults by using the existing model-based techniques, parameters of motor control voltage, motor current, operation time, and rotational speed were collected by using test bench for data-based failure diagnosis. Thus, the normal and abnormal data of train door were collected and trained by using ML software for failure diagnosis.

The main specifications of the EMU door test bench used in this test are same as those of the vehicle being used, and the details of the main specifications of door system and measurement parameters are shown in Table 1.

The failure and fault of the electric door of a railway vehicle differ in the type and frequency of failure depending on the railway operating country. In the case of the Korean metropolitan area, commuting trains cause various door failures due to crowed passengers because so many passengers board and get off. Door failures can be classified according to the level of failure as follows.

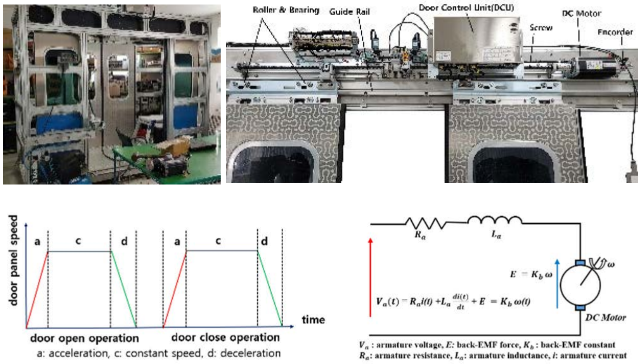

Roller & Bearing
Guide Rail
Door Control Unit(DCU)
DC Motor
Screw
Encorder
door panel speed
a
c
d
a
c
d
time
door open operation
door close operation
a: acceleration, c: constant speed, d: deceleration
R_a
L_a
$E = K_b\,\omega$
ω
$V_a(t) = R_a i(t) + L_a \frac{di(t)}{dt} + E = K_b\,\omega(t)$
DC Motor
V_a : armature voltage, E: back-EMF force, K_b : back-EMF constant
R_a: armature resistance, L_a: armature inductance, i: armature current

FIGURE 1. Railway vehicle door system: test bench; operation equipment; train door operation; train door control.

TABLE 1. The main technical specifications of EMU door, parameters and measurement resolutions of door unit. (sampling time: 50msec).

Equipment	Item/parameters	Specification
electrical pocket sliding door	driving parts	electrical motor, coupler, driving screw, nut, roller, bearing, guide rail
	operating voltage	rating DC 100V
	operating time	2.5/3.0±0.5 sec
	driving screw	Ø18×1,696mm
	door size(W×H)	1,300mm×1,860mm
	stroke	2×650mm=1,300mm
door control unit (DCU)	control voltage	DC 100V
	communication port	MVB, RS485, RS232, USB
motor	type	DC permanent magnet motor with encoder
	no-load rotational speed	1,000 rpm
	current	0.12A(no-load 1,000rpm)
	Torque	0.9Nm/A
	encoder voltage	12VDC
door unit parameter	motor voltage	1V*
	motor current	1A*
	door panel speed	0.001m/s (motor rpm conversion) *

* sampling resolution, sampling time(50msec)

1) Light failure: A frequently occurring failure that includes a fault that causes passengers to overboard during door operation. The types of failures include door panel pushing, opposite direction loading, obstacle entrapment in the door pocket, bearing or roller damage, spindle vibration due to impact, door reopening due to obstruction of passenger or passenger's belongings, and guide rail bending, etc. If the train crew knows the cause of the failure, they can lock the door or take quick emergency measures.
2) Heavy failure: It is a failure that requires replacement or repair of parts after arrival at the base because it is impossible to take action during operation due to damage or failure of major parts. This type of failure includes control power supply cutoff, motor failure, transfer mechanism breakage, and DCU failure. In the case of such a failure, the door panel itself does not operate, so failure data is not collected.

Figure 2 is a block diagram of a door model and a fault simulation for generating data for electrical door fault diagnosis. In this model, failure simulation was performed by selecting 7 representative failures that can obtain door panel movement data. Heavy failure, which does not obtain data because the door panel does not move, is excluded because it cannot be a classification problem by ML. The door model is simple because the motor rotates forward and backward when the DCU sends an operation signal. Therefore, it is composed of spindles, screws, door panels, rollers, and bearings that operate the door. In the fault simulation, when the selected fault load is applied during panel operation, seven types of abnormal data are collected, and normal data are collected during normal operation without a fault load.

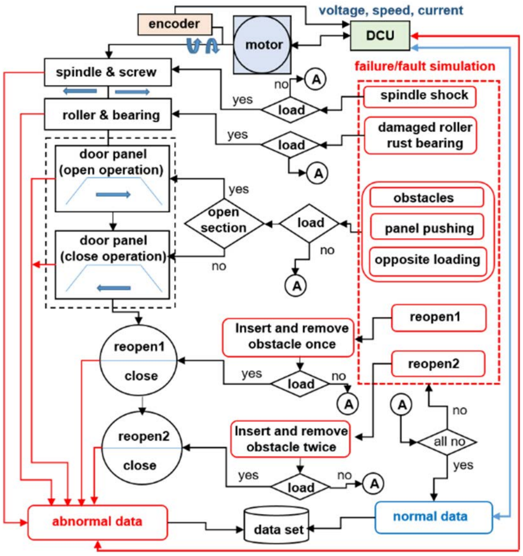


FIGURE 2. Block diagram of a door model and a fault simulation for generating data for electrical door fault diagnosis.

The eight classes for ML door failure diagnosis were selected by analyzing past failure data during the operation of trains in the Seoul metropolitan area by the Korea Railroad Corporation (KORAIL) and the Seoul Metro. At that time, maintenance engineers and door manufacturers were participated. The test was performed after confirming the data by simulating similar to the actual type of failure. The test bench has operating switches that open and close the door, and uses the same DCU as the actual vehicle, the door panels, and the driving mechanism. Voltages, current, and encoder signals are collected and stored through the communication port of the DCU. Data were collected by dividing into open and close operations. The method of assigning loads of failure conditions to obtain abnormal data is as follows.

1) Normal: Perform normal door open/close action
2) Obstacles: Obstacle environment between body sidewall and door panel interferes with normal movement by inserting solid material between roller and rail
3) Door reopen1: To implement an environment that resists movement of door panels during the door close process by the passenger's handbag or other belongings, insert solids between door panels temporarily and remove them to allow one short open close
4) Door reopen 2: Insert and remove solids so that the door reopen is repeated briefly twice
5) Pushing: An adult applies a vertical load to the door panel surface with the palm of the hand to simulate the condition of pushing the door by the passenger during rush hour (currently pushing device is under developing)
6) Spindle shock: Manually and temporarily shake the motor and spindle shaft to simulate abnormal vibration of the transfer mechanism.
7) Opposite loading: Adult obstructing door open by applying palm load in opposite direction of panel moving (JR uses adhesive sheet to panel surfaces)
8) Damaged roller and bearing: Using damaged rollers and rusted bearings to implement transfer resistance

The details of test method of each failure modes and number of datasets are summarized in Table 2. A total of 2,476 dataset were collected with 8 classes of failure mode including normal operation. The normal and abnormal data samples in typical open and close operation are shown in Figure 3.

TABLE 2. Types of failures, test methods and number of measured dataset.

Failure type	Test method and condition	Collected dataset*
normal	normal door open and close operation	625/625
obstacles	insert an obstacle between rollers and rail	50/50
door reopen1	insert and remove an obstacle during the close operation (door reopen once)	50/50
door reopen2	insert and remove an obstacle during the close operation (door reopen twice)	50/50
pushing	apply a vertical load on the door panels to interrupt with operation	100/100
spindle shocks	random shaking on the motor and spindle	50/50
opposite loading	artificial load applied in the opposite direction of door movement	50/50
damaged roller and bearing	installation and use of damaged rollers and rusty bearing	263/263

B. POSITION-BASED DATA CONVERSION

The data segmentation method uses derivatives of encoder speed signal to divide the acceleration, constant speed, and deceleration zone for time domain feature extraction. To divide the data with deviation into uniform zones, methods such as attaching trailing zeros to short data by large number of long data or dividing them into several intervals to obtain an average of the number of data per interval and resampling the data to the same length [27]. In this study, a new method instead of existing method was developed by converting time-based data into distance-based or position-based data. The converted data could be used for feature extraction of three-segmented data without losing parts of data.

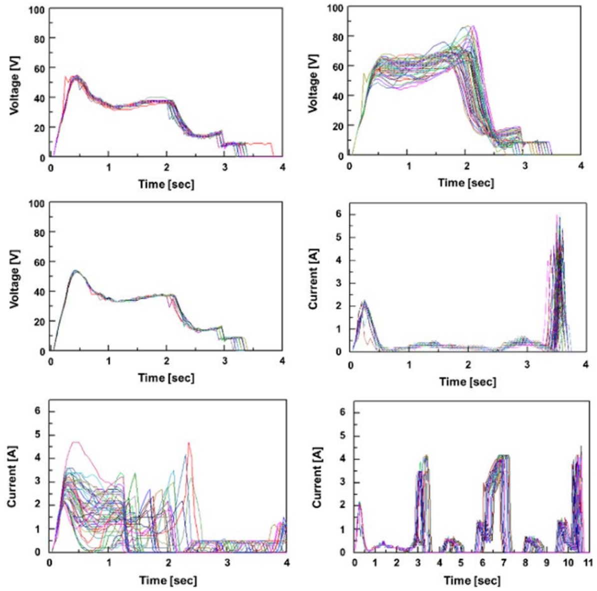


FIGURE 3. Data samples during the door normal and abnormal operation: normal open voltage; opposite loading open voltage; reopen 1 open voltage; normal close current; opposite loading close current; reopen 2 close current.

The time axis data are converted to a position axis value based on the 650mm stroke of the door movement. To convert current or voltage data into position-based data, the travel distance of door panels obtained and can be calculated as follows:

$$d_{op} = \int_{0}^{t_{open}} v_{op}(t)\,dt \tag{1}$$

$$d_{cl} = \int_{0}^{t_{close}} v_{cl}(t)\,dt \tag{2}$$

where, d_{op} is distance, v_{op} is velocity of open operation, d_{cl} is distance, v_{cl} is velocity of close operation, t is time.

The position-based data are obtained by accumulating the distance data as follows:

$$p_{op} = \sum\nolimits_{i=2}^{k_{op}} d_{opi}\,(x_i - x_{i-1}) \tag{3}$$

$$p_{cl} = \sum_{j=2}^{k_{cl}} d_{clj}\,(x_j - x_{j-1}) \tag{4}$$

where, k_{op} is number of open operation data, x_i is open distance of i_{th} data, if $i = 1$, $x_{i-1} \neq 0$ and $x_i = 0$, then $d_{opi+1} = -1 * d_{opi-1}$ and $d_{opi} = 0$. k_{cl} is number of close operation data, x_j is j_{th} data of close distance, if $j = 1$, $x_{j-1} \neq 0$ and $x_j = 0$, then $d_{clj+1} = -1 * d_{clj-1}$ and $d_{clj} = 0$.

The conversion data samples from time-based data to position-based current data are shown in Figure 4.

C. METHOD OF ESTIMATION AN OPTIMIZED SINGLE DATA ZONE

GA is widely used in optimization to identify superior genes through the hybridization of the dominant genes. It is a method used to find solution spaces through evolution simulations by the 'Survival of the fittest' through selection, mutation and crossover. This method optimizes early generations of elite genes in population through breeding until a desired level of excellence occurs.

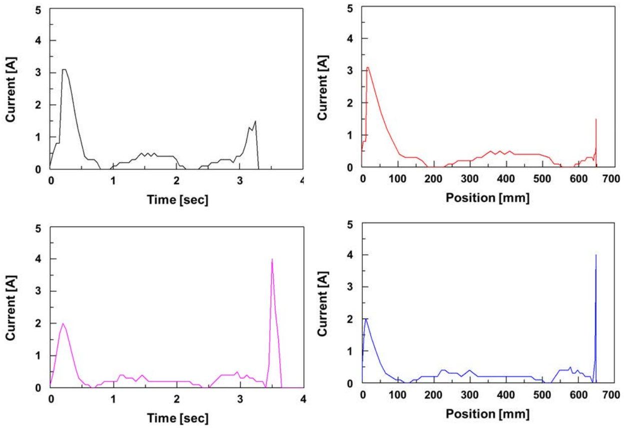


FIGURE 4. Current data samples converted to position x- axis: time-based open current; position-based open current; time-based close current; position-based close current.

To reduce high dimensional features of data, a GA was used to find the optimal data area for feature extraction of each open close operation respectively. The method could produce excellent performance by finding an optimized data zone, starting with the initial full data. This OSDZ can reduce the feature extraction calculations to twice instead of six in the conventional 3 zone segmentation method. The resultant interval values of zone initially performed are used for default data range for feature extraction.

The flowcharts of the OSDZ using a GA are shown in Figure 5. Upper flowchart is a module that sets initial values and sets feature extraction data zone by using PCA and classifier models. This module outputs the classification performance value as the feature value obtained using a specified region through PCA and classification model. Lower flowchart is a GA part module of optimized single zone estimation (OSZE) for feature extraction. The settings used in this processing are as follows: number of genes of population N = 50; permittivity = 50%; max. not improved number = 30; crossover rate = 95% and mutation rate = 5%. The procedure for performing this GA is as follows:

1) Permittivity mutation rate, maximum number of unimproved and data start-end position initial values of x_1, x_2, (open); x_3, x_4 (close) are prepared. The first initial value of data zone is the start and end position of each entire data region.
2) Use single zone data to extract features. PCA-based feature dimensional reduction is performed, and ML is carried out by classification model to obtain performance.
3) The 4 initial position values are reset according to the permittivity and the variation rate, and the data region is

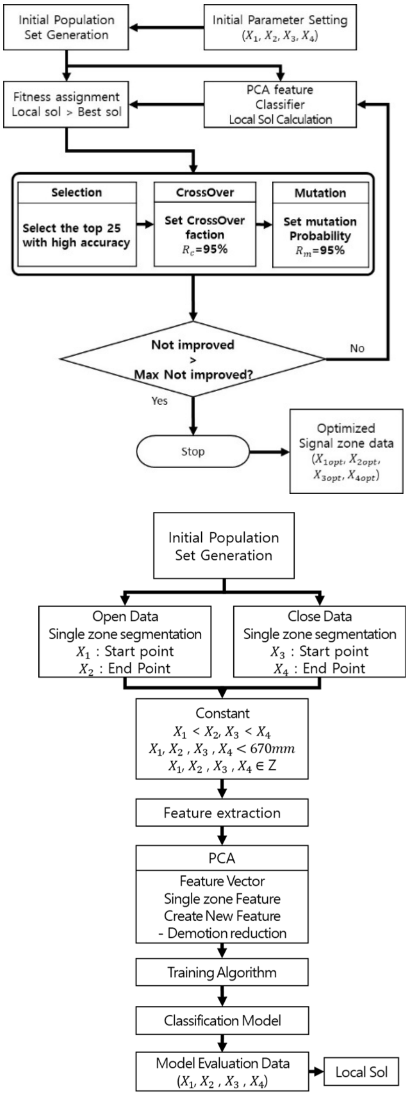


FIGURE 5. Flowchart of the optimized single zone estimation using a GA: flow chart of module of initial parameter setting algorithm including PCA and classifier; GA for optimized single zone estimation for feature extraction.

reduced. Repeat ML, if the classification performance obtained by the new interval data is improved over the previous value, the result is stored as a global solution.

4) Repeat the same process and update the global solution if better performance is found than previous global solution values. If these iterative renewals do not occur up to the maximum number of unimproved, the process is stopped, and the optimal approximation is obtained.

The conventional 3 zone segmentation method divides each parameter into three sections. Thus, a total of 78 statistical features are extracted per parameter by each open and close data, if the statistical 13 features are used. The data section divided by MZS and OSZE for feature extraction are shown in Figure 6. The figure shows that why the feature dimension can be reduced by using OSZE method.

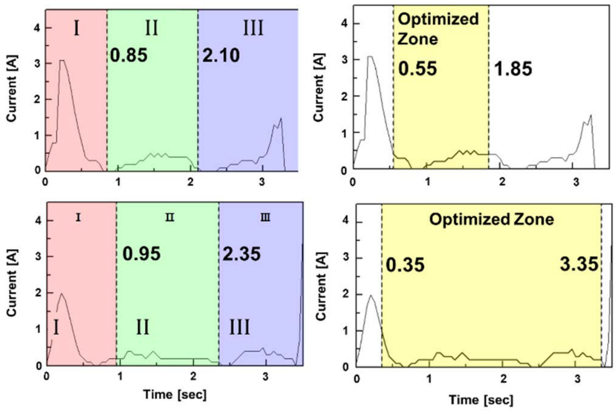


FIGURE 6. Segmentation techniques for feature extraction: 3 zone segmentation for time-based open data; OSZE for position-based open data; 3 zone segmentation for time-based close data; OSZE for position-based close data. (I: acceleration, II: constant speed, III: deceleration).

D. FEATURE SELECTION TECHNIQUES

Feature selection is the linear transformation technique that select the key features that failures can classify well from normal signals, and LDA methods are often used. One of commonly used methods of LDA technique is using FDR and PCC.

1) FISHER'S DISCRIMINANT RATIO

FDR is a measure of how well a single feature classifies two classes of normalized features by mean and standard deviation. The best feature can be selected by obtaining the FDR is obtained [7], [42].

Find a feature f_1 with FDR the best of the features.

$$f_1 = \underset{m}{\operatorname{argmax}}\ FDR \tag{5}$$

Next, the second best, feature f_2 is obtained by the following equation.

$$f_2 = \underset{m}{\operatorname{argmax}}\ \{\omega FDR_m (j, k) - (1 - \omega) \left|\rho_{f_1 m}\right|\} \tag{6}$$

where, f_1 is the ID of the first best feature, m is a feature ID excluding f_1, $\rho_{f_1 m}$ is cross correlation for two features of ids f_1 and m, ω is the weighted value set by relative importance.

2) PEARSON'S CORRELATION COEFFICIENT

From the features of abnormal signals, PCC can be used to select highly correlated features. The correlation ρ_0 and PCC between X and Y consisting of n samples are as follows [31], [43].

$$\boldsymbol{\rho_0} = \frac{\boldsymbol{Cov(X, Y)}}{\sqrt{\boldsymbol{\sigma_X^2 \sigma_Y^2}}} \tag{7}$$

$$\mathbf{PCC\,(X, Y)} = \frac{\boldsymbol{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}}{\sqrt{\boldsymbol{\sum_{i=1}^{n} (x_i - \bar{X})^2 (y_i - \bar{Y})^2}}} \tag{8}$$

where, $\bar{X}$ and $\bar{Y}$ is the mean; σ_X^2 and σ_Y^2 is variance components of variables.

The range of PCC(X,Y) is from -1 to 1. The value of -1 indicates negative correlation, 1 positive correlation, and 0 indicates irrelevant correlation. Typically, the degree of correlation could be divided into 0∼0.4 is weak linear correlation; 0.4∼0.7 significant; and 0.7∼1.0 is strong linear correlation. Using this coefficient, correlation between features can be analyzed to select key features. In this work, PCC(X,Y) > 0.85 is used first before PCA is performed to ensure that the features with strong correlation are selected.

3) PCA-BASED FEATURE DIMENSIONAL REDUCTION

PCA techniques are widely used in unsupervised learning algorithms, which are used for ML modeling as techniques used for dimensional reduction. It is a statistical approach that reduces the high dimensional feature vector set to a new low dimensional vector set. It is a method of normalizing the data first, constructing the covariance matrix, obtaining eigenvectors, and then finding the representative eigenvalues. It is a statistical approach that reduces the high dimensional feature vector set to a new low dimensional vector set. This technique assumes that most of the information in the classes includes variations is the largest.

The procedure for dimensional reduction of the feature vector set $X = [x_1, x_2, x_3, \cdots, x_M]$ of N-dimension, $x_i \in R^{N \times M}$ to low p-sensitive features is as follows: where M is number of features, and $1 \leq p \leq M$ [44], [45].

1) Calculate the mean value.
2) Find the covariance matrix of features.

$$\text{Cov}\ (\text{x}) = \frac{1}{M} \sum\nolimits_{i=1}^{M} (x_i - \mu)^T (x_i - \mu) \tag{9}$$

3) Decompose covariance matrix to obtain eigenvalues and eigenvector, then sorting by descend order. Obtain p dimensional feature subspace (k ≤ M) by computing p largest eigenvalues and corresponding eigenvector.

$$Y = \left[y_1, \cdots, x_p\right] = \left[T_1^T X, \cdots T_p^T X\right] = T^T X \tag{10}$$

The explained variance ratio (EVR) is defined as follows.

$$R_p = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{11}$$

To find the principal component axis, find the first axis with maximum variance in the feature vector set. Find the second axis with maximum variance while orthogonal to the found axis. The third axis is orthogonal to the first and second axes and finds the axis that preserves the variance as much as possible. In the same way, the axis is found by the dimension of the dataset. The flowchart of the PCA algorithm is shown in Figure 7.

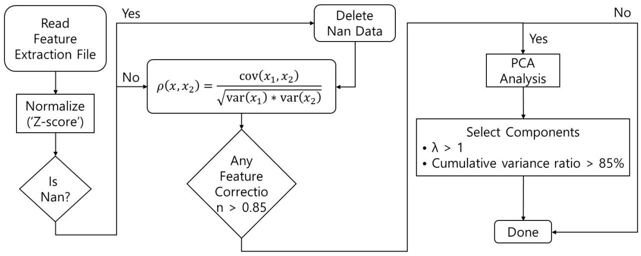

Read Feature Extraction File
Normalize ('Z-score')
Is Nan?
Yes
No
Delete Nan Data
$\rho(x, x_2) = \frac{\text{cov}(x_1, x_2)}{\sqrt{\text{var}(x_1) * \text{var}(x_2)}}$
Any Feature Correctio n > 0.85
Yes
No
PCA Analysis
Select Components
• λ > 1
• Cumulative variance ratio > 85%
Done

FIGURE 7. Flowchart of the PCA algorithm.

E. FAILURE DIAGNOSIS CLASSIFIER AND PERFORMANCE INDICATOR

Six representative classification models were used to obtain the classification performance of feature sets. The classifiers, kNN, SVM, decision tree, Naïve Bayes, and ensemble are already librarised, so they can be easily coded and used. The classification accuracy is compared using door data labeled by failure category. When diagnosing motor failure using current, voltage, etc. variables, the SVM classifier is often used due to its high accuracy [33]. In addition, the kNN classifier is often used for diagnosing rotating machine failures due to simple implementation and clear performance [46].

Datasets were allocated and used at 70% for training and 30% for evaluation, respectively. MATLAB, a general-purpose program, was used for coding including proposed algorithms. The classification models used are shown in Table 3. The classifiers used are 6 SVMs; 5 kNN; 5 decision trees; 2 Naïve Bayes; and 1 ensemble with 3 models, respectively [47], [48], [49].

TABLE 3. The classifiers used for ML.

Method	Classifier Model
SVM	linear, quadratic, cubic, fine Gaussian, medium Gaussian, coarse Gaussian
kNN	fine, medium, cosine, weighted, subspace
Decision Trees	fine, medium, boosted, bagged, rusboosted
Naïve Bayes	Gaussian, kernel
Ensemble	Subspace, rusboosted, bagged trees

Indicators evaluating the performance of classification model for ML are accuracy and F1 score, and others defined in Table 4. The confusion matrix shows correct predictions by diagonal lines and incorrect predictive types and describes the complete performance of the model [50].

F. COMPARISON STUDY OF ML-BASED DOOR FAILURE DIAGNOSIS TECHNIQUES

ML-based fault diagnosis is generally performed in four stages: data collection and preprocessing; feature extraction; feature selection; and classification by ML. Feature extraction is that the process of converting and projecting the

TABLE 4. Model performance evaluate metrics.

Indicator	Definition	Formula*
Accuracy	the ratio of correct predictions to the input samples	$\frac{TP+TN}{TP+FP+TN+FN}$
F1 Score	a weighted average of precision and recall	$\frac{2*Precision*Recall}{Precsion+Recall}$
Recall	a measure of a classifiers completeness, the fraction of relevant instances that were retrieved.	$\frac{TP}{TP+FN}$
Precision	the fraction of relevant instances among the retrieved instances	$\frac{TP}{TP+FP}$
Specificity	the proportion of actual negatives	$\frac{TN}{FP+TN}$

*TP = TruePositives; TN = TrueNegatives; FP = FalsePositives; FN = FaultNegatives.

collected data into a new low dimensional feature space without losing its nature of data.

The techniques proposed in this work in the overall flow of ML for door failure diagnosis is illustrated in Figure 8. In the figure, the blue dotted line is a feature extraction and feature selection technique, including the collection and preprocessing of data used in conventional door failure diagnosis techniques. The flow marked by red dotted lines is the method proposed in this study. It is a method of rearranging data based on distance, estimating an OSDZ using a GA, and applying PCA algorithms. To compare the proposed techniques with conventional methods, 4 cases were examined by feature dimension as shown in Table 5.

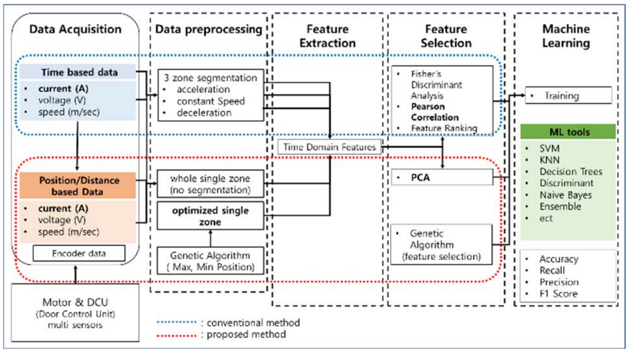


FIGURE 8. Research framework of RVDS failure diagnosis: data preprocessing, feature extraction, feature selection, and ML.

The details of parameter, data segmentation techniques used, and total features extracted, in each case are described as follows:

1) Case 1: Position-based current data are used. Features are extracted from all data zone data. 26 features;
2) Case 2: Position-based current data are used. Features are extracted from OSDZ using a GA, and PCA-based feature deduction method. 26 features;
3) Case 3: Time-based current data are used. Features are extracted from the acceleration, constant speed, and deceleration sections of the three segmented data. 78 features;
4) Case 4: Position-based current, voltage, speed data are used. Features are extracted from the acceleration, constant speed, and deceleration sections of the three-zone segmented data. 234 features.

TABLE 5. Feature dimensions by case.

Case	Parameters	Data conversion	Data segmentation*	Methods	Features extracted(p)
1	current	position-based	1 zone (whole data)	PCA	2×1×13×1 (26)
2	current	position-based	1 zone (OSDZ)	GA and PCA	2×1×13×1 (26)
3	current	time-based	3 zone (acc, con, dec)	FDA and PCC	2×1×13×3 (78)
4	current, voltage, speed (3)	position-based	3 zone (acc, con, dec)	FDA and PCC	2×3×13×3 (234)

*acc: acceleration; con: constant speed; dec: deceleration.

To compare the preprocessing effects of data, case 1 to 2, and case 4 used position-based data and cases 3 used time-based data. To compare the variable selection effect, case 1-3 used current data, while case 4 used voltage, current and speed data. In addition, to compare the effects of data segmentation techniques, case 1 and 2 used full data and OSDZ respectively, while case 3 and case 4 used 3 zone segmented data.

In this work, 13 representative statistical features were used as described in Table 6.

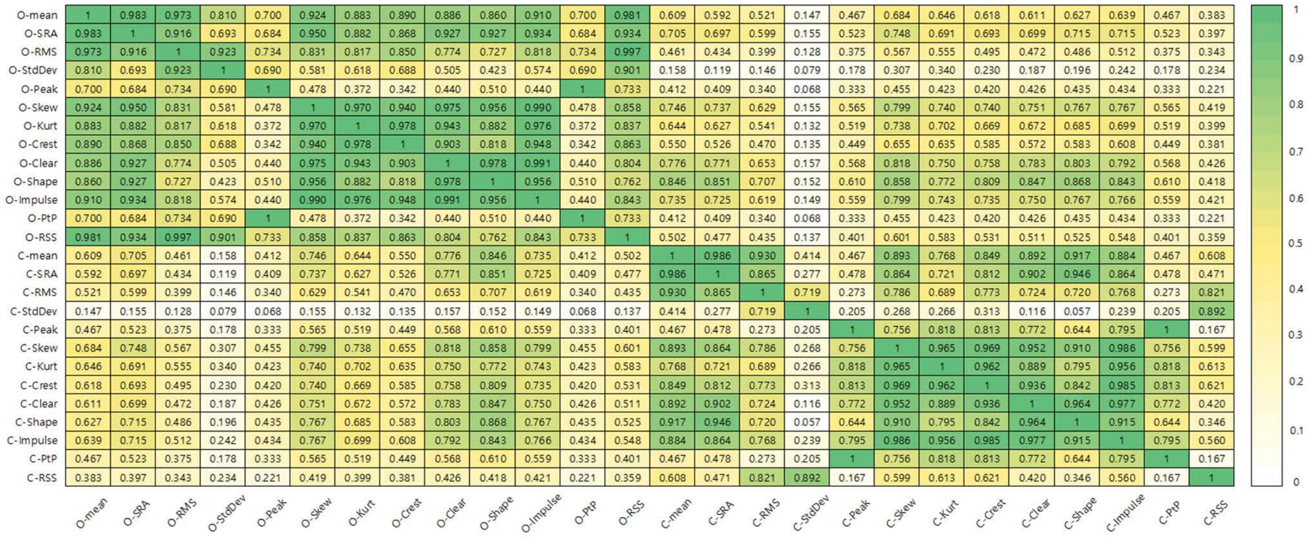
III. RESULTS

A. PCA-BASED FEATURE DIMENSIONAL REDUCTION: CASE 1

Measured data with several parameters increases the dimension of features. The multi-segmentation of data leads to higher feature dimension. Therefore, there is a probability of high correlation between individual features, which requires a more intuitive interpretation of the data by dimensional reduction. The multi-collinearity was investigated by analyzing the correlation of a total of 26 features of case 1 in Table 5.

The heat map plot of correlation of 26 features by PCC is shown in Figure 9. Correlation analysis shows the degree of association between features, the correlation coefficient is PCC with absolute value is used. The x-axis represents 13 characteristics of the open operation, and the y-axis represents 13 features of the close operation.

Table 7 shows a high and low correlation between open and close operation. These highly correlated features cause a multi-collinearity problem. To avoid this, it is necessary to remove them by feature selection.



	O-mean	O-SRA	O-RMS	O-StdDev	O-Peak	O-Skew	O-Kurt	O-Crest	O-Clear	O-Shape	O-Impulse	O-PtP	O-RSS	C-mean	C-SRA	C-RMS	C-StdDev	C-Peak	C-Skew	C-Kurt	C-Crest	C-Clear	C-Shape	C-Impulse	C-PtP	C-RSS
O-mean	1	0.983	0.973	0.810	0.700	0.924	0.883	0.890	0.886	0.860	0.910	0.700	0.981	0.609	0.592	0.521	0.147	0.467	0.684	0.646	0.618	0.611	0.627	0.639	0.467	0.383
O-SRA	0.983	1	0.916	0.693	0.684	0.950	0.882	0.868	0.927	0.927	0.934	0.684	0.934	0.705	0.697	0.599	0.155	0.523	0.748	0.691	0.693	0.699	0.715	0.715	0.523	0.397
O-RMS	0.973	0.916	1	0.923	0.734	0.831	0.817	0.850	0.774	0.727	0.818	0.734	0.997	0.461	0.434	0.399	0.128	0.375	0.567	0.555	0.495	0.472	0.486	0.512	0.375	0.343
O-StdDev	0.810	0.693	0.923	1	0.690	0.581	0.618	0.688	0.505	0.423	0.574	0.690	0.901	0.158	0.119	0.146	0.079	0.178	0.307	0.340	0.230	0.187	0.196	0.242	0.178	0.234
O-Peak	0.700	0.684	0.734	0.690	1	0.478	0.372	0.342	0.440	0.510	0.440	1	0.733	0.412	0.409	0.340	0.068	0.333	0.455	0.423	0.420	0.426	0.435	0.434	0.333	0.221
O-Skew	0.924	0.950	0.831	0.581	0.478	1	0.970	0.940	0.975	0.956	0.990	0.478	0.858	0.746	0.737	0.629	0.155	0.565	0.799	0.740	0.740	0.751	0.767	0.767	0.565	0.419
O-Kurt	0.883	0.882	0.817	0.618	0.372	0.970	1	0.978	0.943	0.882	0.976	0.372	0.837	0.644	0.627	0.541	0.132	0.519	0.738	0.702	0.669	0.672	0.685	0.699	0.519	0.399
O-Crest	0.890	0.868	0.850	0.688	0.342	0.940	0.978	1	0.903	0.818	0.948	0.342	0.863	0.550	0.526	0.470	0.135	0.449	0.655	0.635	0.585	0.572	0.583	0.608	0.449	0.381
O-Clear	0.886	0.927	0.774	0.505	0.440	0.975	0.943	0.903	1	0.978	0.991	0.440	0.804	0.776	0.771	0.653	0.157	0.568	0.818	0.750	0.758	0.783	0.803	0.792	0.568	0.426
O-Shape	0.860	0.927	0.727	0.423	0.510	0.956	0.882	0.818	0.978	1	0.956	0.510	0.762	0.846	0.851	0.707	0.152	0.610	0.858	0.772	0.809	0.847	0.868	0.843	0.610	0.418
O-Impulse	0.910	0.934	0.818	0.574	0.440	0.990	0.976	0.948	0.991	0.956	1	0.440	0.843	0.735	0.725	0.619	0.149	0.559	0.799	0.743	0.735	0.750	0.767	0.766	0.559	0.421
O-PtP	0.700	0.684	0.734	0.690	1	0.478	0.372	0.342	0.440	0.510	0.440	1	0.733	0.412	0.409	0.340	0.068	0.333	0.455	0.423	0.420	0.426	0.435	0.434	0.333	0.221
O-RSS	0.981	0.934	0.997	0.901	0.733	0.858	0.837	0.863	0.804	0.762	0.843	0.733	1	0.502	0.477	0.435	0.137	0.401	0.601	0.583	0.531	0.511	0.525	0.548	0.401	0.359
C-mean	0.609	0.705	0.461	0.158	0.412	0.746	0.644	0.550	0.776	0.846	0.735	0.412	0.502	1	0.986	0.930	0.414	0.467	0.893	0.768	0.849	0.892	0.917	0.884	0.467	0.608
C-SRA	0.592	0.697	0.434	0.119	0.409	0.737	0.627	0.526	0.771	0.851	0.725	0.409	0.477	0.986	1	0.865	0.277	0.478	0.864	0.721	0.812	0.902	0.946	0.864	0.478	0.471
C-RMS	0.521	0.599	0.399	0.146	0.340	0.629	0.541	0.470	0.653	0.707	0.619	0.340	0.435	0.930	0.865	1	0.719	0.273	0.786	0.689	0.773	0.724	0.720	0.768	0.273	0.821
C-StdDev	0.147	0.155	0.128	0.079	0.068	0.155	0.132	0.135	0.157	0.152	0.149	0.068	0.137	0.414	0.277	0.719	1	0.205	0.268	0.266	0.313	0.116	0.057	0.239	0.205	0.892
C-Peak	0.467	0.523	0.375	0.178	0.333	0.565	0.519	0.449	0.568	0.610	0.559	0.333	0.401	0.467	0.478	0.273	0.205	1	0.756	0.818	0.813	0.772	0.644	0.795	1	0.167
C-Skew	0.684	0.748	0.567	0.307	0.455	0.799	0.738	0.655	0.818	0.858	0.799	0.455	0.601	0.893	0.864	0.786	0.268	0.756	1	0.965	0.969	0.952	0.910	0.986	0.756	0.599
C-Kurt	0.646	0.691	0.555	0.340	0.423	0.740	0.702	0.635	0.750	0.772	0.743	0.423	0.583	0.768	0.721	0.689	0.266	0.818	0.965	1	0.962	0.889	0.795	0.956	0.818	0.613
C-Crest	0.618	0.693	0.495	0.230	0.420	0.740	0.669	0.585	0.758	0.809	0.735	0.420	0.531	0.849	0.812	0.773	0.313	0.813	0.969	0.962	1	0.936	0.842	0.985	0.813	0.621
C-Clear	0.611	0.699	0.472	0.187	0.426	0.751	0.672	0.572	0.783	0.847	0.750	0.426	0.511	0.892	0.902	0.724	0.116	0.772	0.952	0.889	0.936	1	0.964	0.977	0.772	0.420
C-Shape	0.627	0.715	0.486	0.196	0.435	0.767	0.685	0.583	0.803	0.868	0.767	0.435	0.525	0.917	0.946	0.720	0.057	0.644	0.910	0.795	0.842	0.964	1	0.915	0.644	0.346
C-Impulse	0.639	0.715	0.512	0.242	0.434	0.767	0.699	0.608	0.792	0.843	0.766	0.434	0.548	0.884	0.864	0.768	0.239	0.795	0.986	0.956	0.985	0.977	0.915	1	0.795	0.560
C-PtP	0.467	0.523	0.375	0.178	0.333	0.565	0.519	0.449	0.568	0.610	0.559	0.333	0.401	0.467	0.478	0.273	0.205	1	0.756	0.818	0.813	0.772	0.644	0.795	1	0.167
C-RSS	0.383	0.397	0.343	0.234	0.221	0.419	0.399	0.381	0.426	0.418	0.421	0.221	0.359	0.608	0.471	0.821	0.892	0.167	0.599	0.613	0.621	0.420	0.346	0.560	0.167	1

FIGURE 9. Correlation analysis of 26 statistical features of open-close current data: case 1 (o-: open, c-: close).

TABLE 6. Statistical feature formulas extracted with N data points.

Feature ID	Feature	Equation
1	mean	$F_m = \frac{1}{N}\sum_{n=1}^{N} x(n)$
2	square root amplitude	$F_{sra} = (\frac{1}{N}\sum_{n=1}^{N}\sqrt{\lvert x(n)\rvert})^2$
3	root mean square	$F_{rms} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2}$
4	standard deviation	$F_{std} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n)-F_m)^2}$
5	peak	$F_p = max(\lvert x(n)\rvert)$
6	skewness	$F_{sk} = \frac{1}{N}\sum_{n=1}^{N}(\frac{x(n)-F_m}{F_{std}})^3$
7	kurtosis	$F_k = \frac{1}{N}\sum_{n=1}^{N}(\frac{x(n)-F_m}{F_{std}})^4$
8	crest factor	$F_{crf} = \frac{F_p}{F_{rms}}$
9	clearance factor	$F_{clf} = \frac{F_p}{\left(\frac{\sum_{n=1}^{N}\sqrt{\lvert x(n)\rvert}}{N}\right)^2}$
10	shape factor	$F_{sf} = \frac{F_{rms}}{F_v}$
11	impulse factor	$F_{if} = \frac{F_p}{F_v}$
12	peak-to-peak	$F_{pp} = max(x(n)) - min(x(n))$
13	root sum square	$F_{rss} = \sqrt{\sum_{n=1}^{N} x(n)^2}$

TABLE 7. Results of correlation analysis of features of door operation.

Relation	High correlation	Low correlation
open-open	F_m - F_{sra}, F_{if} - F_{clf}, F_{clf} - F_{sf}, F_{if} - F_{crf}	F_{pp} - F_{rms}, F_{std} - F_m
open-close	F_{sf} - F_{sra}, F_{pp} - F_{sra}, F_{sf} - F_{sra}, F_{clf} - F_m	F_{sk} - F_{sra}, F_{clf} - F_k
close-close	F_{clf} - F_{sf}, F_m - F_{sra}, F_p - F_{std}, F_{pp} - F_{std}	F_{if} - F_{rms}, F_m - F_k

To reduce the effect of multi-collinearity problem, PCA-based feature dimensional reduction technique was applied. In this work, 4 principal component axes are used and EVR of 98% was found. Figure 10 represent the relationship of 4 principal components and EVR. The y-axis represents the EVR and 98% points are shown by dotted line.

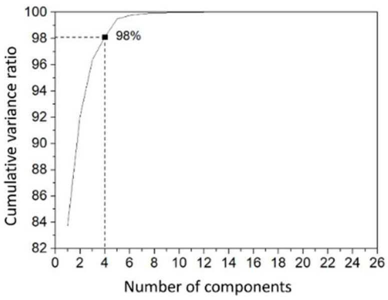


FIGURE 10. The relationship of 4 principal components selection.

Figure 11 shows the correlation ratio of open-close 26 features of 4 principal components. A total of five classes of 20 classification models (SVM, kNN, decision trees, Naïve Bayes, discriminant) were used to compare performance evaluation metrics, with the following results as shown in Table 8.

1) Classifier such as linear, quadratic and cubic SVM show the highest accuracy of 99.52%.

2) F1 score was high at 99.19% linear SVM, 99.14% quadratic and 99.13% cubic SVM.

SVM and kNN classifiers have shown more than 99% accuracy and have high classification performance.

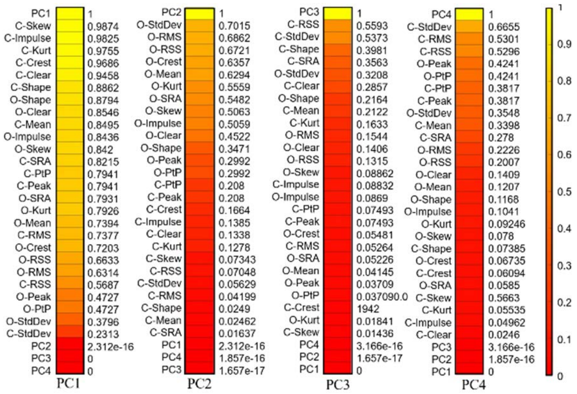


FIGURE 11. Covariance matrix of open-close 26 features of PCA.

TABLE 8. Performance evaluation metrics after PCA: Case 1.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)
linear SVM	99.52	99.09	99.29	99.19
quadratic SVM	99.52	99.21	99.07	99.14
cubic SVM	99.52	98.98	99.29	99.13
fine kNN	99.29	98.94	98.45	98.70
medium kNN	99.29	98.71	98.85	98.78
bagged trees	99.37	98.71	98.71	98.71
Gaussian Naïve Bayes	98.65	97.23	97.52	97.38
kernel Naïve Bayes	99.05	98.14	98.09	98.12
subspace discriminant	97.31	94.86	97.50	96.16

B. OSDZ USING A GA AND FEATURE SELECTION BY PCA: CASE 2

Since there exist difficulties of segmentation to distinguish characteristics of the failure data, a technique to optimize by the classification performance of feature extraction division by removing them has been tested. This technique is a method that reduces the data size to be extracted from the entire data region by applying a metaheuristic technique of GA with position-based current values. After feature extraction using an optimal single section of each open close data, a feature selection technique that minimizes the number of characteristic dimensions by applying the PCA-based feature dimensional reduction method was used.

Fig. 12 shows the feature extraction zone of each open close data optimized for classification performance using a GA. The optimized zone of open data was selected from 110mm to 481mm, and that of close data was from 40mm to 635mm in the 650mm stroke distance of the door.

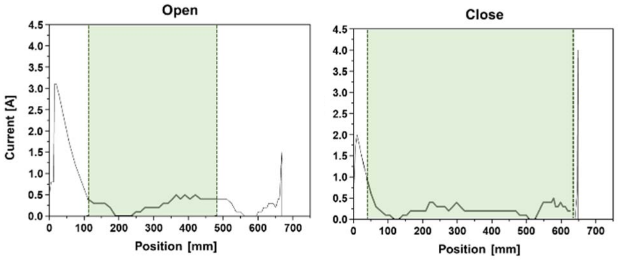
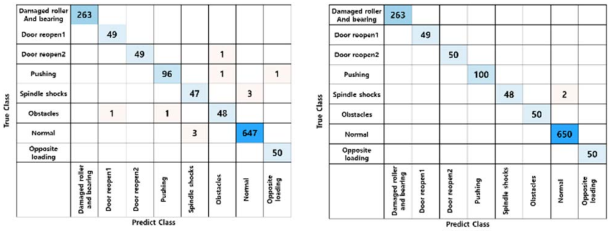


FIGURE 12. Estimated optimized single-zone using a GA: open obstacle data; close obstacle data.

Comparing the data zones for feature extraction of open and close operation, the data range of close operation is relatively wide. This seems to be due to more characteristic's changes in close operations than to open the failure type data experimented on the door test bench. It is estimated that the test conducted for data collection was due to fewer failures in the deceleration section near the end of the open operation, and the frequency of failures in passengers or obstacle simulation tests in the end of close operation was high. Although affected by acquired failure data, this method uses single minimum interval data, which has the advantage of reducing feature extraction time for big data analysis. Therefore, using a GA reduces the data interval used for feature extraction rather than using the entire data, reducing the computational time.

The results of the diagnosis of feature extraction from OSDZ using a GA and feature selection by PCA are shown in Figure 13. It can be seen that the case of using the OSDZ is better classification performance than the case of using the entire data.


True Class
Predict Class
True Class
Predict Class

	Damaged roller and bearing	Door reopen1	Door reopen2	Pushing	Spindle shocks	Obstacles	Normal	Opposite loading
Damaged roller And bearing	263							
Door reopen1		49						
Door reopen2			49			1		
Pushing				96		1		1
Spindle shocks					47		3	
Obstacles		1		1		48		
Normal					3		647	
Opposite loading								50

	Damaged roller and bearing	Door reopen1	Door reopen2	Pushing	Spindle shocks	Obstacles	Normal	Opposite loading
Damaged roller And bearing	263							
Door reopen1		49						
Door reopen2			50					
Pushing				100				
Spindle shocks					48		2	
Obstacles						50		
Normal							650	
Opposite loading								50

FIGURE 13. Results of the diagnosis of feature extraction from optimized data zone using a GA and feature selection by PCA: confusion matrix using PCA from full data; confusion matrix using PCA from optimization zone data.

Table 9 represents the changes in data zones and the classification accuracy with generational changes in GA processing. GA is set to terminate if no improvements of up to 20 generations are made. Finally, over a total of 26 generations, the accuracy improved from 99.37% to 99.84%.

The classification performance of application the GA and PCA algorithm are shown Table 10. After carrying out ML classification, the highest accuracy of 99.84% of cubic SVM was shown.

TABLE 9. Optimization for feature extraction data zone and classification accuracy improvement by GA.

Generation number	Data range		Accuracy (%)
	Open	Close	
1	full range	full range	99.37
3	224 - 449	28 - 620	99.52
4	110 - 515	30 - 536	99.60
5	26 - 578	27 - 549	99.68
16	279 - 444	7 - 613	99.76
26	95 - 471	15 - 619	99.84

TABLE 10. Classification performance of application of a GA and PCA algorithm: Case 2.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)
linear SVM	99.68	99.34	99.29	99.19
cubic SVM	99.84	99.82	99.29	99.13
medium Gaussian SVM	99.60	99.46	98.45	98.70
cubic kNN	99.21	98.48	98.85	98.78
weighted kNN	99.21	98.59	98.71	98.71
bagged Trees	98.21	98.46	97.52	97.38
Gaussian Naïve Bayes	98.34	96.87	98.09	98.12
subspace discriminant	96.20	91.36	97.50	96.16

C. FEATURE SELECTION BY FDR AND PCR: CASE 3

Feature selection was performed by recent study using FDR and PCR, for comparison with the proposed methods. First, calculate the FDR value for each feature of 28 combinations of 8 categories of door failures, and the highest FDR is selected as the best feature. The second-best feature is selected as a feature that has a high FDR but has a minimum correlation to avoid duplicate selection. (See Reference 7 for detailed instructions) As a result of the FDR selection, the open operating features are ranked 4,11,7,6,9,3,13. The feature selection results of the close operation are ranked 13, 4, 11, 6, 9. Where, the number is the feature id number of Table 11 and 12 are the results of feature selection using FDR and PCC. The results are similar to those of previous studies.

The classification performances of feature selection using FDR and PCC are shown in Table 13. The results are linear SVM, fine kNN, and quadratic SVM shows 99.21% of high accuracy. Fine kNN shows 98.22%, quadratic SVM 98.21%, and linear SVM 98.15% of F1 scores.

D. FEATURE SELECTION BY FDR AND PCR: CASE 4

In this case, feature selection was performed by FDR and PCR with 234 features in the 3 segmented zones of acceleration, constant speed, deceleration and the position-based data of 3 parameters of current, voltage and speed.

The classification performances of feature selection using the methods are shown in Table 14. Quadratic SVM show 99.60% of high accuracy and 99.27% of F1 score.

TABLE 11. Feature selection using FDR and PCR, time-based current data with 3 zone data of open operation: Case 3.

OPEN OPERATION					
Class order (Combination)	Feature ID		Class order (COMBINATION)	Feature ID	
	Best	2nd		Best	2nd
Class 1&2	11	7	Class 3&5	10	9
Class 1&3	4	9	Class 3&6	4	13
Class 1&4	11	4	Class 3&7	6	13
Class 1&5	4	11	Class 3&8	5	7
Class 1&6	4	13	Class 4&5	9	11
Class 1&7	7	3	Class 4&6	4	1
Class 1&8	6	7	Class 4&7	11	6
Class 2&3	4	9	Class 4&8	6	7
Class 2&4	11	5	Class 5&6	4	2
Class 2&5	11	3	Class 5&7	4	1
Class 2&6	4	3	Class 5&8	6	7
Class 2&7	11	5	Class 6&7	4	13
Class 2&8	3	9	Class 6&8	3	9
Class 3&4	6	13	Class 7&8	5	7

TABLE 12. Feature selection using FDR and PCR, time-based current data with 3 zone data of close operation: Case 3.

CLOSE OPERATION					
Class order (Combination)	Feature ID		Class order (COMBINATION)	Feature ID	
	Best	2nd		Best	2nd
Class 1&2	9	7	Class 3&5	13	1
Class 1&3	4	9	Class 3&6	10	5
Class 1&4	13	11	Class 3&7	13	11
Class 1&5	13	6	Class 3&8	4	6
Class 1&6	13	7	Class 4&5	13	9
Class 1&7	13	1	Class 4&6	13	4
Class 1&8	4	6	Class 4&7	13	9
Class 2&3	11	3	Class 4&8	13	4
Class 2&4	4	2	Class 5&6	8	11
Class 2&5	11	7	Class 5&7	4	11
Class 2&6	1	9	Class 5&8	13	6
Class 2&7	11	5	Class 6&7	6	1
Class 2&8	11	5	Class 6&8	10	13
Class 3&4	13	4	Class 7&8	13	6

TABLE 13. The classification performance of feature selection technique of FDR and PCC: Case3.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)
linear SVM	99.21	98.23	98.08	98.15
quadratic SVM	99.21	98.11	98.31	98.21
fine kNN	99.21	97.98	98.45	98.22
bagged trees	98.89	97.33	97.91	97.62
Gaussian Naïve Bayes	98.50	96.30	97.23	96.76
kernel Naïve Bayes	96.52	94.82	93.67	94.24

E. COMPARISON OF FAULT DIAGNOSIS PERFORMANCE BY FEATURE DESIGN CASES

The results of comparing the feature design techniques proposed in chapter 2 with the existing methods are shown in Table 15 (see Table 6) and Figure 14.

TABLE 14. The classification performance of feature selection technique of FDR and PCC: Case4.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)
linear SVM	99.13	98.66	98.46	98.56
quadratic SVM	99.60	99.31	99.23	99.27
fine kNN	99.21	98.72	98.56	98.64
bagged trees	99.21	98.76	98.62	98.69
Gaussian Naïve Bayes	98.81	97.74	98.00	97.87
kernel Naïve Bayes	98.81	97.94	98.70	98.32

TABLE 15. Diagnosis performance of classification models by feature design case.

Applied techniques	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)
1 zone full data and PCA(current)	99.52	98.98	99.29%	99.13
GA optimized 1 zone and PCA(current)	99.84	99.82	99.29	99.13
3 zone segmentation and FDA(current)	99.21	98.11	98.31	98.21
3 zone and FDA(current, voltage, speed)	99.60	99.31	99.23	99.27

1) Case 1: 26 features, which uses position-based current data and used 13 statistical features in the whole data zone shows high accuracy of 99.52% and F1 score of 99.13%.
2) Case 2: 26 features, which uses position-based current data and feature extraction applied OSZE using a GA and PCA-based feature reduction shows the best performance of the cubic SVM model with 99.84% accuracy and 99.13% F1 score.
3) Case 3: 78 features, which obtained by dividing time-based current data into 3 sections, and applied FDA and PCR also shows relatively high classification performance with 99.21% accuracy and 98.21% F1 score.
4) Case 4: 234 features, which obtained in the 3 segmented zones of acceleration, constant speed, deceleration zones of position-based data of current, voltage and speed, shows 99.60 % of accuracy and 99.27% of the highest F1 score.

F. CASE OF FDA APPLICATION AFTER FEATURE EXTRACTION IN GA OPTIMIZED SECTION

The optimization section of position-based current data was determined using GA for each open and closed data, and 13 statistical features were obtained. Then, the diagnostic performance was obtained after feature selection by applying the existing FDA technique.

Table 16 shows the results of diagnostic performance in this case. The quadratic SVM showed the highest accuracy of 99.60% and the F1 score of 99.22%.

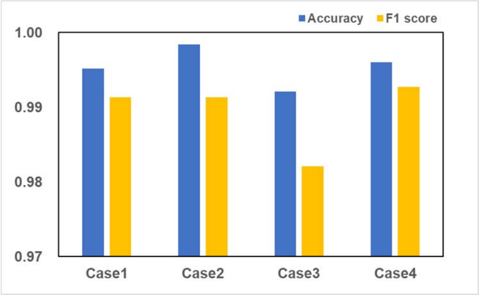


FIGURE 14. The classification performance indicators by feature design cases: accuracy; F1 score.

TABLE 16. The classification performance of Feature selections by current whole data and pca: fda-ga(fg) case.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)
quadratic SVM	99.60	99.09	99.36	99.22
fine KNN	99.52	99.20	99.41	99.31
bagged trees	99.13	98.38	98.18	98.28
RUS boosted trees	98.81	97.32	97.77	97.54
Gaussian Naïve Bayes	98.42	97.41	96.05	96.73
kernel Naïve Bayes	97.54	95.36	95.92	95.64

G. CASE OF USING 3 PARAMETERS OF TIME-BASED DATA

For comparison of diagnostic performance when using time-based data and position-based data, diagnostic performance was obtained using FDA method with three parameter data: current, voltage, and speed of measured time-based data.

Table 17 shows the results of diagnostic performance. The bagged trees showed the highest accuracy of 94.30% and the F1 score value of 88.18%.

TABLE 17. The classification performance of Feature selections by 3 variables with 3 zone(time-based): time-based 3 zone fda(tb3f) case.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)
quadratic SVM	93.58	87.45	84.40	85.90
fine KNN	91.60	85.36	85.11	85.24
bagged trees	94.30	90.09	86.35	88.18
RUS boosted trees	85.50	81.88	89.98	85.74
Gaussian Naïve Bayes	85.18	82.49	89.10	85.67
kernel Naïve Bayes	84.55	79.84	88.35	83.88

H. COMPARISON OF DIAGNOSTIC PERFORMANCE WITH CONVENTIONAL FDA METHOD AND TIME-BASED DATA

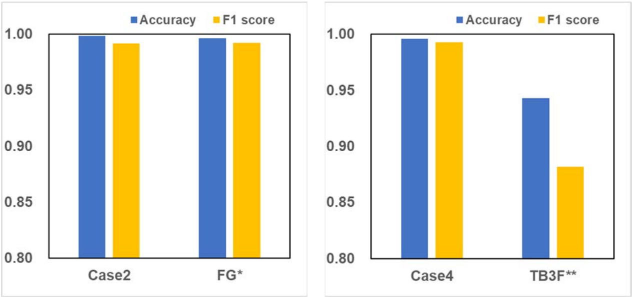
Finally, the effects of the two techniques were compared.

First, the case 2 of applying the PCA technique and the case of applying the FDA for feature dimensional reduction were compared. In both cases, position-based current data were optimized with GA and statistical features were extracted.

Second, the effects of time-based data and position-based data were compared. In both cases, three voltage, current, and speed parameters were used. In consideration of the three variables, feature extraction was performed after dividing into three sections, and the FDA technique was applied to reduce the feature dimension.

Figure 15 shows a comparison of these techniques. The first figure is a graph comparing the accuracy of using conventional FDA techniques with Case 2 in which PCA was applied to reduce feature dimensions. The second figure is a graph comparing the diagnostic accuracy when using time-based data and position-based data. The comparison results are as follows:

1) In the case of the feature dimension reduction effect, it was found that dimension reduction by PCA application rather than FDA application was effective when one current parameter was used. (case 2 and FG case)
2) As for the data conversion effect, it was found that position-based data conversion (case 4) was more effective when all three variables were used. (case 4 and TB3F case)



*FG: FDA-GA; **TB3F: Time-based 3 zone FDA

FIGURE 15. The classification performance comparison: case2 and FG case; case4 and TB3F case.

I. PRACTICAL APPLICABILITY OF PROPOSED TECHNIQUES

For practical door failure diagnosis, GA and PCA techniques were proposed as feature dimensional reduction methods, and higher classification performance was obtained while using lower dimension features than previous studies.

1) The GA method can minimize the feature extraction data to the optimal section. Therefore, it is possible to significantly reduce the feature extraction time of the data in the process of preprocessing the vast amount of data generated by many doors in real time.
2) The accuracy of fault diagnosis classification could be improved even by using three to four features by the PCA technique.
3) In this way, GA and PCA techniques can be used to process actual vehicle data in the future because they can obtain high classification accuracy while using small data. In particular, the GA method can be used while upgrading the optimization section after a certain period without running every time unless the characteristics of each door change significantly due to maintenance.
4) The PCA technique can be used to select and use only the minimum features that increase classification accuracy, so it can be useful for actual door failure diagnosis.

IV. DISCUSSION

Converting time-based data into position-based data has an effect on improving the classification performance. This alternative method of data conversion can be used without the data being discarded for feature extraction.

The performance of using a GA and PCA-based feature dimensional reduction method of case 2 and case 4, both show the best performance of 99.84% accuracy. Therefore, the current value data can be used for a parameter for fast and effective failure diagnosis of RVDS. The PCA technique selected four main components, resulting in a higher diagnostic accuracy than the recent methods of feature selection using FDR and PCR. It was confirmed that the PCA-based feature dimensional reduction could improve the diagnosis performance than the existing technique of feature selection. In addition, since the OSZE using a GA can be performed once to change the feature extraction zone when fault data are accumulated in a period of a certain train operation, only the set minimum data is used for quick diagnosis. Among various classification models for ML, SVM classifiers such as quadratic SVM and cubic SVM had the highest classification performance for RVDS failure diagnosis. To improve the classification performance in supervised learning, sufficient quantities of abnormal data should be obtained. In this experiment, a relatively large number of data could be collected from the test bench to achieve high accuracy.

In the future, a comparative study of this study will be conducted when the door data collection device is installed in the actual vehicles. In addition, future research related to forecasting maintenance cycles and RULs are planned to be carried out by analyzing the time-series data of major components e.g. bearings and motors.

V. CONCLUSION

Combination of the optimization of feature extraction zones using a GA and PCA-based feature dimensional reduction method with converted 8 classes of signals acquired using the door test bench were proposed. With this practical feature design techniques, various comparative studies were conducted to improve fault and failure classification performance. The conclusions obtained through the studies by ML of test data are as follows:

1) Using the combination of all proposed methods with the converted position-based current data, the feature extraction using a GA and the PCA-based feature dimensional reduction, shows the best accuracy of 99.84%.

2) In terms of feature extraction technique, the classification performance was improved if features were extracted once from the OSDZ using a GA rather than the existing multi-segmented method. The existing 3-zone segmentation method was 99.21% of accuracy, which was lower to 99.52% when the entire current data were used without segmentation (case 1). Because case of feature extraction from the entire data without data segmentation can obtain higher accuracy, it was confirmed that this method was available considering the computational time.
3) A high accuracy of 99.60% was achieved in case of using all parameter of voltage, current, and speed using FDA with 234 features (case 4). However, in case of time-based data, 94.30% of accuracy was relatively low when position-based conversion was not made.

In conclusion, the combination of the aforementioned methods will be practical to enable fast diagnosis and to ensure classification accuracy of failure diagnosis from data collected simultaneously at many doors of railway vehicles operated in real commercial lines.

REFERENCES

[1] Z. M. Çinar, A. A. Nuhu, Q. Zeeshan, O. Korhan, M. Asmael, and B. Safaei, "Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0," *Sustainability*, vol. 12, no. 19, p. 8211, Oct. 2020.

[2] E. Miguelanez, K. E. Brown, R. Lewis, C. Roberts, and D. M. Lane, "Fault diagnosis of a train door system based on semantic knowledge representation," in *Proc. 4th IET Int. Conf. Railway Condition Monitor. (RCM)*, 2008, pp. 1–6.

[3] C. Cheng, J. Wang, H. Chen, Z. Chen, H. Luo, and P. Xie, "A review of intelligent fault diagnosis for high-speed trains: Qualitative approaches," *Entropy*, vol. 23, no. 1, p. 1, Dec. 2020.

[4] N. Lehrasab, H. P. B. Dassanayake, C. Roberts, S. Fararooy, and C. J. Goodman, "Industrial fault diagnosis: Pneumatic train door case study," *Proc. Inst. Mech. Eng., F, J. Rail Rapid Transit*, vol. 216, no. 3, pp. 175–183, May 2002.

[5] Y. Sun, G. Xie, Y. Cao, and T. Wen, "Strategy for fault diagnosis on train plug doors using audio sensors," *Sensors*, vol. 19, no. 1, p. 3, Dec. 2018.

[6] Y. Cao, Y. Sun, G. Xie, and T. Wen, "Fault diagnosis of train plug door based on a hybrid criterion for IMFs selection and fractional wavelet package energy entropy," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7544–7551, Aug. 2019.

[7] S. Ham, S.-Y. Han, S. Kim, H. J. Park, K.-J. Park, and J.-H. Choi, "A comparative study of fault diagnosis for train door system: Traditional versus deep learning approaches," *Sensors*, vol. 19, no. 23, p. 5160, Nov. 2019.

[8] W. Shi, N. Lu, B. Jiang, Y. Zhi, and Z. Xu, "Incipient fault diagnosis method of railway vehicle door system based on random forest," in *Proc. Chin. Control Conf. (CCC)*, Jul. 2019, pp. 4901–4906.

[9] Y. H. Wang, L. F. Bi, and L. J. Li, "Reliability analysis of plug door system based on dynamic fault tree," *Appl. Mech. Mater.*, vol. 627, pp. 207–211, Sep. 2014.

[10] Q. L. Zhou, B. Y. Jin, and Z. Y. Xing, "Reliability analysis and fault diagnosis of metro door system based on Bayesian network," *J. Shenyang Univ. Technol.*, vol. 36, no. 4, pp. 441–445, 2014.

[11] R. Chen, S. Zhu, F. Hao, B. Zhu, Z. Zhao, and Y. Xu, "Railway vehicle door fault diagnosis method with Bayesian network," in *Proc. 4th Int. Conf. Control Robot. Eng. (ICCRE)*, Apr. 2019, pp. 70–74.

[12] O. Fink, E. Zio, and U. Weidmann, "Fuzzy classification with restricted Boltzman machines and echo-state networks for predicting potential railway door system failures," *IEEE Trans. Rel.*, vol. 64, no. 3, pp. 861–868, Sep. 2015.

[13] F. Fang, Z.-J. Zhao, C. Huang, X.-Y. Zhang, H.-T. Wang, and Y.-J. Yang, "Application of reliability-centered maintenance in metro door system," *IEEE Access*, vol. 7, pp. 186167–186174, 2019.

[14] Z. Liming, C. Guoqiang, Y. Jianwei, and J. Limin, "Monte-Carlo simulation based on FTA in reliability analysis of door system," in *Proc. 2nd Int. Conf. Comput. Autom. Eng. (ICCAE)*, Feb. 2010,, pp. 713–717.

[15] X. Cheng, Z. Xing, Y. Qin, Y. Zhang, S. Pang, and J. Xia, "Reliability analysis of metro door system based on FMECA," *J. Intell. Learn. Syst. Appl.*, vol. 5, no. 4, pp. 216–220, 2013.

[16] H. Dassanayake, C. Roberts, C. J. Goodman, and A. M. Tobias, "Use of parameter estimation for the detection and diagnosis of faults on electric train door systems," *Proc. Inst. Mech. Eng., O, J. Risk Rel.*, vol. 223, no. 4, pp. 271–278, Oct. 2009.

[17] L. Cauffriez, S. Grondel, P. Loslever, and C. Aubrun, "Bond graph modeling for fault detection and isolation of a train door mechatronic system," *Control Eng. Pract.*, vol. 49, pp. 212–224, Apr. 2016.

[18] A. Boussif and M. Ghazel, "Model-based monitoring of a train passenger access system," *IEEE Access*, vol. 6, pp. 41619–41632, 2018.

[19] P. Fraga-Lamas, T. Fernández-Caramés, and L. Castedo, "Towards the internet of smart trains: A review on industrial IoT-connected railways," *Sensors*, vol. 17, no. 6, p. 1457, Jun. 2017.

[20] T. Böhm, "Remaining useful life prediction for railway switch engines using classification techniques," *Int. J. Prognostics Health Manag.*, vol. 8, no. 3, pp. 1–15, Nov. 2020.

[21] W. Mao, J. He, J. Tang, and Y. Li, "Predicting remaining useful life of rolling bearings based on deep feature representation and long short-term memory neural network," *Adv. Mech. Eng.*, vol. 10, no. 12, Dec. 2018, Art. no. 168781401881718.

[22] H. Lee, S.-Y. Han, and K.-J. Park, "Generative adversarial network-based missing data handling and remaining useful life estimation for smart train control and monitoring systems," *J. Adv. Transp.*, vol. 2020, pp. 1–15, Nov. 2020.

[23] X. Tao, C. Ren, Q. Li, W. Guo, R. Liu, Q. He, and J. Zou, "Bearing defect diagnosis based on semi-supervised kernel local Fisher discriminant analysis using pseudo labels," *ISA Trans.*, vol. 110, pp. 394–412, Apr. 2021.

[24] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[25] A. Mangal and E. A. Holm, "A comparative study of feature selection methods for stress hotspot classification in materials," *Integrating Mater. Manuf. Innov.*, vol. 7, no. 3, pp. 87–95, Jun. 2018.

[26] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Amer. Statist.*, vol. 42, no. 1, pp. 59–66, 1988.

[27] A. W.-C. Fu, E. Keogh, L. Y. H. Lau, C. A. Ratanamahatana, and R. C.-W. Wong, "Scaling and time warping in time series querying," *VLDB J.*, vol. 17, no. 4, pp. 899–921, Mar. 2007.

[28] P.-F. Marteau, "Time warp edit distance with stiffness adjustment for time series matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 306–318, Feb. 2009.

[29] J. Fortuna and D. Capson, "Improved support vector classification using PCA and ICA feature space modification," *Pattern Recognit.*, vol. 37, no. 6, pp. 1117–1129, Jun. 2004.

[30] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, Sep. 1933.

[31] F. R. S. K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philosoph. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.

[32] L. B. Jack and A. K. Nandi, "Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms," *Mech. Syst. Signal Process.*, vol. 16, nos. 2-3, pp. 373–390, Mar. 2002.

[33] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, Aug. 2007.

[34] X. Zhu, J. Xiong, and Q. Liang, "Fault diagnosis of rotation machinery based on support vector machine optimized by quantum genetic algorithm," *IEEE Access*, vol. 6, pp. 33583–33588, 2018.

[35] Z. Chen, T. Lin, N. Tang, and X. Xia, "A parallel genetic algorithm based feature selection and parameter optimization for support vector machine," *Sci. Program.*, vol. 2016, no. 2, pp. 1–10, Jun. 2016.

[36] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning K for kNN classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, pp. 1–19, 2017.

[37] L. Kou, Y. Qin, and X. Zhao, "An integrated model of kNN and GBDT for fault diagnosis of wheel on railway vehicle," in *Proc. Prognostics Syst. Health Manag. Conf. (PHM-Chongqing)*, Oct. 2018, pp. 432–436.

[38] L. Cauffriez, P. Loslever, N. Caouder, F. Turgis, and R. Copin, "Robustness study and reliability growth based on exploratory design of experiments and statistical analysis: A case study using a train door test bench," *Int. J. Adv. Manuf. Technol.*, vol. 66, nos. 1–4, pp. 27–44, Jul. 2012.

[39] L. Cauffriez, R. Copin, N. Caouder, P. Loslever, and F. Turgis, "Design of a testing bench for simulating tightened-up operating conditions of train's passenger access," in *Reliability, Risk, and Safety*. Boca Raton, FL, USA: CRC Press, Aug. 2009.

[40] T. Akaogi, J. Mishima, T. Ichigi, and Y. Sugiura, "Study on failure sign detection using monitoring data for door operating equipment of commuter trains," *JR EAST Tech. Rev.*, vol. 29, pp. 26–29, Apr. 2014.

[41] T. Isao, U. Kotaro, and F. Kenji, "Passenger door system for series E235 train of east Japan railway company (yamanote line) designed to improve transportation quality," *Fuji Electric Rev.*, vol. 64, no. 1, pp. 44–48, 2018.

[42] S. Wang, D. Li, Y. Wei, and H. Li, "A feature selection method based on Fisher's discriminant ratio for text sentiment classification," in *Proc. Web Inf. Syst. Mining*. Berlin, Germany: Springer, 2009, pp. 88–97.

[43] J. Yan and J. Lee, "Degradation assessment and fault modes classification using logistic regression," *J. Manuf. Sci. Eng.*, vol. 127, no. 4, pp. 912–914, Jul. 2004.

[44] Y. Li, W. Dai, and W. Zhang, "Bearing fault feature selection method based on weighted multidimensional feature fusion," *IEEE Access*, vol. 8, pp. 19008–19025, 2020.

[45] J. N. R. Jeffers, "Two case studies in the application of principal component analysis," *Appl. Statist.*, vol. 16, no. 3, pp. 225–236, 1967.

[46] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Feb. 2018.

[47] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, "A dual coordinate descent method for large-scale linear SVM," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 408–415.

[48] P. A. Lachenbruch and M. Goldstein, "Discriminant analysis," *Biometrics*, vol. 35, no. 1, pp. 69–85, 1979.

[49] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May/Jun. 1991.

[50] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *Proc. Australas. Joint Conf. Artif. Intell.* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2006, pp. 1015–1021.

GIL HYUN KANG was born in South Korea, in 1958. He received the B.S. degree in chemical mechanical engineering from Chonnam National University, Gwangju-si, in 1981, and the M.phil. and Ph.D. degrees in mechanical engineering from the University of Birmingham, Birmingham, U.K., in 2011. From 1981 to 2007, he was at the Headquarter of Rolling Stock Division and Workshops in Korean National Railroad. He was the Executive Director and the President of World Railway Research Congress, from 2003 to 2007. He was at the Research Center and Rolling Stock Business, Group of Hyundai Rotem Company, and SR High Speed Company, from 2010 to 2017. Since 2017, he has been a Research Dedicated Professor with the Industry-Academic Cooperation Foundation, Korea National University of Transportation, Uiwang, South Korea. He has authored of two books, more than ten articles, and holds two patents. His research interest includes smart maintenance of rolling stock. He is a Reviewer of the *International Journal of Railways* in Korean Society for Railways.

HO CHEOL KI received the B.S. and M.S. degrees in mechanical engineering from Sungkyunkwan University, in 2003. From 2003 to 2005, he was a Research Assistant with the High-Speed Train Engineering Corps, Korea Railroad Research Institute. He was at the Sound and Vibration Team of Cylos Company and Hyundai-Rotem Company, from 2006 to 2017. Since 2017, he has been Researcher with the Industry-Academic Cooperation Foundation, Korea National University of Transportation, Uiwang, South Korea. His research interest includes development of fault diagnosis technology for railway vehicle parts using artificial intelligence.

SANG HYUN AN received the B.S. degree in industrial engineering from Ajou University, Suwon, South Korea, in 2021, where he is currently pursuing the M.S. degree in industrial engineering. From 2013 to 2019, he was a Software Engineer at the Global Technical Center, Samsung Electro-Mechanics, Suwon. From 2020 to 2021, he was a Software Engineer at the Department of AI, GS Solutions, Uiwang, South Korea. His research interests include the development of abnormal detection and fault diagnosis using machine learning method, reinforcement learning, modeling and simulation, image processing, and development of image processing and motion control techniques using machine learning method.

JUHEE CHOI (Member, IEEE) received the B.S. degree in computer science from Yonsei University, in 2004, and the M.S. and Ph.D. degrees in computer science and engineering from Seoul National University, in 2006 and 2016, respectively. From 2006 to 2020, he was at the SoC Development Team, Samsung Electronics Company Ltd. He is currently an Assistant Professor with the Department of Smart Information and Communication Engineering, Sangmyung University. His research interests include the IoT systems, railway safety, and railway diagnosis systems.

CHUL SU KIM was born in South Korea, in 1972. He received the B.S. degree in mechanical engineering from Hanyang University, Seoul, in 1996, where he received the M.S. and Ph.D. degrees in mechanical design, in 2002. Since 2003, he has been a Professor with the Department of Railway Engineering, Korea National Transportation University, Uiwang, Gyeonggi. From 2013 to 2017, he was a member of a Management Evaluation Committee for local public corporations, Ministry of Safety and Public Administration of Korea. From 2014 to 2015, he was a Visiting Researcher at Korea Railroad Corporation. He has authored three books, more than 50 articles, and holds ten patents. His research interest includes smart maintenance of rolling stock. He is a Reviewer of the *International Journal of Railway* in Korean Society for Railways.