applied sciences

MDPI

Article

A Study on PF–IFF-Based Diagnosis Model of Plant Equipment Failure

Min-Young Seo [1], Se-Yun Hwang [2], Jang-Hyun Lee [3], Jae-Gon Kim [4] and Hong-Bae Jun [5,*]

1 EENTA Co., Seongnam 13449, Korea; minyoung819.seo@eenta.co.kr
2 Extreme Technology Research Center for Ship and Offshore Platform, INHA University, Incheon 22212, Korea; syhwang@inha.ac.kr
3 Department of Naval Architecture and Ocean Engineering, INHA University, Incheon 22212, Korea; jh_lee@inha.ac.kr
4 Department of Industrial and Management Engineering, Incheon National University 119 Academy, Yeonsu-gu, Incheon 22012, Korea; jaegkim@inu.ac.kr
5 Department of Industrial Engineering, Hongik University, 94 Wausan-ro, Mapo-gu, Seoul 04066, Korea
* Correspondence: hongbae.jun@hongik.ac.kr; Tel.: +82-2-320-3056

Abstract: There are two types of maintenance policies for equipment: breakdown maintenance and preventive maintenance. In the case of applying preventive maintenance, the maintenance is carried out based on time or the condition of the equipment. However, with the development of Information and Communications Technologies (ICT) and the Internet of Things (IoT) technology, the data collected from equipment has rapidly increased and the use of Condition-Based Maintenance (CBM) to perform appropriate maintenance based on the condition of the equipment is increasing. In this study, based on gathered sensor data, we introduce an approach to diagnosing the condition of the equipment by extracting specific data features related to the types of failures that occur with equipment. To this end, we used the K-means clustering method, support vector machine (SVM) classifier, and Pattern Frequency–Inverse Failure mode Frequency (PF–IFF) method with the Term Frequency–Inverse Document Frequency (TF–IDF) method. As a case study, we applied the proposed approach to a centrifugal pump and carried out computational experiments for assessing the performance and validity of the proposed approach.

Keywords: CBM; diagnosis; SVM; TF–IDF; plant equipment failure

Citation: Seo, M.-Y.; Hwang, S.-Y.; Lee, J.-H.; Kim, J.-G.; Jun, H.-B. A Study on PF–IFF-Based Diagnosis Model of Plant Equipment Failure. *Appl. Sci.* **2022**, *12*, 347. https://doi.org/10.3390/app12010347

Academic Editor: Jordi Cusido

Received: 22 November 2021
Accepted: 28 December 2021
Published: 30 December 2021

1. Introduction

In recent years, emerging technologies such as Information and Communications Technologies (ICT) and Industrial Internet of Things (IIoT) allow us to collect and utilize the data of the status of equipment of major plant facilities and equipment. A plant is a very complex system with several years of design and development followed by several decades of operation and maintenance, consequently, producing a huge amount of data during its life cycle. In particular, with the plant operation data, it is important to keep plant equipment intact during the long service life of the plant. This is because small failures or breakdowns of plant equipment can cause great losses throughout the plant's operations.

For example, in chemical plants such as petroleum refineries, an accident due to an equipment failure such as leakage can bring catastrophic damage not only to industrial workers but also to the local community. Additionally, power plants such as nuclear power and thermal power plants can cause massive losses due to continuous explosions and fire in the event of a disaster due to the continuous processing of high temperature and high pressure and the centralization of utility facilities. Hence, developing an advanced approach for diagnosing the status of plant equipment using collected data to prevent such damage in advance is needed.

In general, there are two kinds of maintenance policy: Breakdown Maintenance (BM) and Preventive Maintenance (PM). BM starts maintenance activities only after a failure

occurs, whereas PM conducts maintenance activities based on time or equipment condition as a preventive measure before a failure occurs. PM can be further divided into Time-Based Maintenance (TBM) and Condition-Based Maintenance (CBM). Considering the risk, the CBM is appropriate for the plant equipment that can cause significant damage to the environment in the event of a failure [1].

This study introduces an equipment diagnosis approach for identifying failure modes with gathered sensor data from plant equipment under the CBM policy. The proposed approach diagnoses what kinds of failures will appear in the future by extracting specific patterns associated with equipment failure modes based on gathered sensor data and establishing diagnosis models using them. In this study, the Term Frequency–Inverse Document Frequency (TF–IDF) technique, mainly used in text mining, is applied to efficiently detect anomalous patterns associated with equipment failure modes. The TF–IDF value tells us how critical a word is in a specific document. Many research works use it as a method for evaluating the degree of importance of words existing in the document. In this study, we used the TF–IDF technique to identify the patterns related to a specific failure mode among the patterns that occurred before the failure mode, and also applied the K-means clustering method and support vector machine (SVM) classifier to establish the diagnosis model.

This study is organized as follows: the relevant previous works are reviewed in Section 2 and the proposed approach is described in Section 3. The case study carried out is discussed in Section 4. Finally, Section 5 concludes the study with a summary.

2. Previous Works

This section addresses relevant previous works: CBM studies, studies using TF–IDF in text mining, and failure diagnosis research works.

2.1. CBM Study

The concept of CBM was first introduced by the Rio Grande Railway Company in the late 1940s and was initially called predictive maintenance [2]. So far, several studies related to CBM methods have been conducted. For example, Bunks et al. (2000) [3] introduced a CBM method using the Hidden Markov Model (HMM) and carried out case studies based on the data collected from the Westland helicopter gearbox. Djurdjanovic et al. (2003) [4] presented the concept of the Watchdog Agent, which is a framework that can perform CBM on machinery using status data gathered by sensor and wireless internet technologies. Jardine et al. (2006) [5] reviewed the CBM research works related to the failure diagnostics and prognostics of mechanical systems, mainly focusing on models, algorithms, and techniques for data processing and maintenance decision-making. Shin and Jun (2015) [2] introduced CBM definitions, related international standards, CBM procedures, and detailed techniques that were dealt with in previous CBM studies and also introduced related case studies. Table 1 summarizes the relevant works.

Table 1. Previous works on CBM.

Authors	Subject	Approach
Bunks et al. (2000) [3]	They introduced a CBM method using the Hidden Markov Model (HMM) and carried out case studies based on the data collected from the Westland helicopter gearbox.	Hidden Markov Model
Djurdjanovic et al. (2003) [4]	Using sensors to collect multiple attributes and wireless internet technology, they proposed the Watchdog Agent concept to carry out CBM on mechanical devices.	Wavelet Transform, Fourier Transform, Autoregressive Moving Average, etc.
Jardine et al. (2006) [5]	They reviewed the CBM research works related to failure diagnostics and prognostics of mechanical systems, mainly focusing on models, algorithms, and techniques for data processing and maintenance decision-making.	CBM review study
Shin and Jun (2015) [2]	They introduced CBM definitions, related international standards, CBM procedures, and detailed techniques that were dealt with in previous CBM studies, and also introduced related case studies.	CBM review study

2.2. TF–IDF Related Studies

The TF–IDF method [6] was mainly used to extract keywords in information retrieval and text mining. It calculates the TF–IDF value to determine how important a word is in a particular document. The following Equations (1)–(3) are formulas for calculating TF–IDF value:

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d) : w \in d\}} \quad (1)$$

$$idf(t,D) = \log(1 + \frac{|D|}{|\{d \in D : t \in d\}|}) \quad (2)$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \quad (3)$$

Here, t indicates a specific word, d indicates a specific document and D indicates a set of documents. w indicates a word included in a specific document and $f(t,d)$ denotes the total frequency of words in the document. $\max\{f(w,d) : w \in d\}$ denotes the frequency of the highest word in document d, $|\{d \in D : t \in d\}|$ denotes the number of documents containing the word t, and $|D|$ indicates the total number of documents. $tf(t,d)$ is a value indicating how frequently a particular word appears in a particular document and $idf(t,D)$ is a value indicating how common a particular word has appeared throughout the document set. $tfidf(t,d,D)$ has a high value when a certain word occurs frequently and only in a specific document. Therefore, the TF–IDF value could be interpreted as a value telling us how important a word is in a particular document.

The previous studies using TF–IDF are as follows. For example, Phelan et al. (2009) [7] proposed a method for recommending news articles using real-time Twitter data and used the TF–IDF technique to rank news. Go et al. (2011) [8] proposed a methodology that enables content-based retrieval using text mining by improving the existing methods of keywords and search expression in the patent search field. Furthermore, Trstenjak et al. (2014) [9] proposed a framework for classifying documents using the k-nearest neighbor algorithm and the TF–IDF technique. Yoo et al. (2015) [10] proposed a method of extracting the main word of novel text using the TF–IDF technique considering the external structure of the novel. Table 2 summarizes the relevant works.

Table 2. Previous works on TF–IDF.

Authors	Subject	Approach
Phelan et al. (2009) [7]	They proposed a method for recommending news articles using real-time Twitter data. When ranking news, they used the TF–IDF technique.	TF–IDF
Go et al. (2011) [8]	They proposed a method that enables content-based retrieval using text mining by improving the existing keywords and search expressions in the field of patent searches. In their approach, they used the TF–IDF technique to extract keywords from documents.	TF–IDF
Trstenjak et al. (2014) [9]	Based on the KNN algorithm and TF–IDF technique, they suggested a document classification framework.	k-nearest neighbor algorithm, TF–IDF
Yoo et al. (2015) [10]	Using the TF–IDF technique and external structure of a novel, they proposed a method of extracting subject words of the novel.	TF–IDF

2.3. Failure Diagnosis Studies

So far, many research works have been proposed for diagnosing equipment conditions and failures. For example, Dong and He (2007) [11] used the Hidden Semi-Markov Model (HSMM) method where the state transition probability changes with time. The proposed method could diagnose the state of the equipment based on the gathered sensor data. Wang and Chen (2007) [12] applied a partly-linearized neural network in order to diagnose centrifugal pump failure. To this end, they extracted relevant features using a continuous wavelet transform technique. Yang and Su (2008) [13] proposed a methodology for diagnosing faults in sewer pipes by extracting organizational characteristics of pipe faults in CCTV images using SVM. Fei and Zhang (2009) [14] applied the SVM with the

Genetic Algorithm (SVMG) method for diagnosing power transformation failures based on gas information gathered from the equipment. Furthermore, Li and Nilkitsaranont (2009) [15] proposed a methodology based on the combination of linear regression and quadratic regression to predict the deterioration state of a gas turbine engine. Son et al. (2009) [16] proposed a Fault Detection System (FDS) using a feature-based decision tree based on the data collected from the sensor of the equipment. Zhou et al. (2010) [17] proposed a methodology for diagnosing equipment failure after extracting signals using the wavelet transform method for vibration data. Wu et al. (2013) [18] proposed a method for estimating the residual life prediction error using an artificial neural network and estimating the failure rate threshold corresponding to the lowest maintenance cost based on the estimation. Muralidharan et al. (2014) [19] performed a study to extract features from a centrifugal pump using a continuous wavelet transform technique and to determine equipment failures with SVM for diagnosing equipment failures with vibration data. Moon and Kim (2014) [20] proposed a diagnostic system using artificial neural networks and wavelet transformations to efficiently diagnose mechanical failures such as mass imbalance and the misalignment of axes that occur frequently in wind power generation systems. Kim et al. (2016) [21] proposed a method in order to estimate the remaining useful life of the pump tower, the main structure of LNG FPSO, using marine environmental data. Recently, Mostafa et al. (2018) [22] proposed an agent-based inference engine and a multi-agent system collaborative module for car failure diagnosis. In their work, they claimed that the failure diagnosis operation is a heuristic and complex series of activities that require prior knowledge and experience regarding the diagnosed object. They addressed domain knowledge, data acquisition methods, detection rules for detecting abnormality, assessment methods for identifying inconsistencies, and inference methods as being a requirement for the automated diagnosis method. Contreras-Valdes et al. (2020) [23] reviewed predictive data mining methods for fault detection and diagnosis in electrical equipment. As for the predictive data mining methods, they introduced the classification-based, regression-based, and hybrid techniques, and reviewed recent research trends on fault diagnosis methods. Abid et al. (2021) [24] reviewed the research works related to Fault Detection and Diagnosis (FDD) methods. In their work, they presented basic FDD system elements with relevant techniques, and also addressed challenges in FDD with future research trends. They mentioned that FDD methods could be grouped into traditional model-based methods and signal processing-based FDD methods with artificial intelligence.

Although there have been many studies on the methods of failure diagnosis of equipment, in this study, we first introduce an approach to identify the state of the equipment with the TF–IDF technique. In many cases of failure diagnosis studies, the wavelet transform technique is mainly used to detect anomalous patterns from univariate (mainly vibration) data. However, in general, for checking equipment status, it is more practical to include more sensor data such as vibration, temperature, pressure, and so on. Hence, this study deals with the failure diagnosis method considering multivariate sensor data. In this study, with multivariable (vibration, temperature, etc.) data, TF–IDF and clustering techniques are used to define the characteristics of the equipment's main state, and then extract the equipment abnormality patterns and use them for equipment state diagnosis. These points mentioned above can be said to be different from existing studies. Table 3 summarizes the relevant works.

Table 3. Previous works on failure diagnosis.

Authors	Subject	Approach
Dong and He (2007) [11]	Unlike HMMs with fixed state transition probability, the HSMM technique considers the change of the state transition probability over time, suggesting a method to diagnose and predict the condition of equipment.	Hidden semi-Markov model
Wang and Chen (2007) [12]	They applied a partly-linearized neural network in order to diagnose the centrifugal pump failure. To this end, they extracted relevant features using a continuous wavelet transform technique.	Continuous Wavelet Transform, Partially-linearized neural network
Yang and Su (2008) [13]	They proposed a methodology to diagnose faults in sewer pipes by extracting organizational characteristics of pipe faults in CCTV images using SVM.	SVM, Radial Basis Network, Back-Propagation Neural network
Fei and Zhang (2009) [14]	They applied the SVM with the Genetic Algorithm (SVMG) method for diagnosing power transformation failures based on gas information gathered from the equipment.	Support Vector Machine with Genetic Algorithm
Li and Nilkitsaranont (2009) [15]	With the model combining linear regression with quadratic regression, they predicted the deterioration of a gas turbine engine.	Regression
Son et al. (2009) [16]	They proposed the FDS (Fault Detection System) using a feature-based decision tree based on the data collected from the sensor of the equipment.	Decision Tree
Wu et al. (2013) [18]	With the estimator of the remaining life prediction error using an artificial neural network, they develop a method estimating the failure rate threshold corresponding to the lowest maintenance cost.	Artificial Neural Network
Muralidharan et al. (2014) [19]	They performed a study to extract features from a centrifugal pump using a continuous wavelet transform technique and to determine equipment failures with SVM for diagnosing equipment failures with vibration data.	Continuous Wavelet Transform, SVM
Moon and Kim (2014) [20]	They proposed a diagnostic system using artificial neural networks and wavelet transformations to efficiently diagnose mechanical failures such as mass imbalance and misalignment of axes that occur frequently in wind power generation systems.	Artificial Neural Network, Wavelet Transform
Kim et al. (2016) [21]	Using the marine environmental data, they suggested a method for estimating the remaining life of a pump tower, the main structure of LNG FPSO.	K-means clustering, Autoregressive Integrated Moving Average
Mostafa et al. (2018) [22]	They proposed an agent-based inference engine and a multi-agent system collaborative module for car failure diagnosis.	Agent-based inference engine based on a forward chaining algorithm
Contreras-Valdes et al. (2020) [23]	They reviewed predictive data mining methods for fault detection and diagnosis in electrical equipment.	Review study
Abid et al. (2021) [24]	They reviewed the research works related to Fault Detection and Diagnosis (FDD) methods.	Review study

3. Approach

This section describes an approach to identify the condition of plant equipment based on gathered sensor data. Figure 1 depicts an overview of the approach proposed in this study.

As shown in Figure 1, the approach we proposed consisted of two parts. One part is the learning process needed to build up the learning model for identifying the state of the plant equipment based on the gathered data and the other part is the prediction process used to foresee the state of the equipment with the learned model. The following is a description of the symbols used in this study.

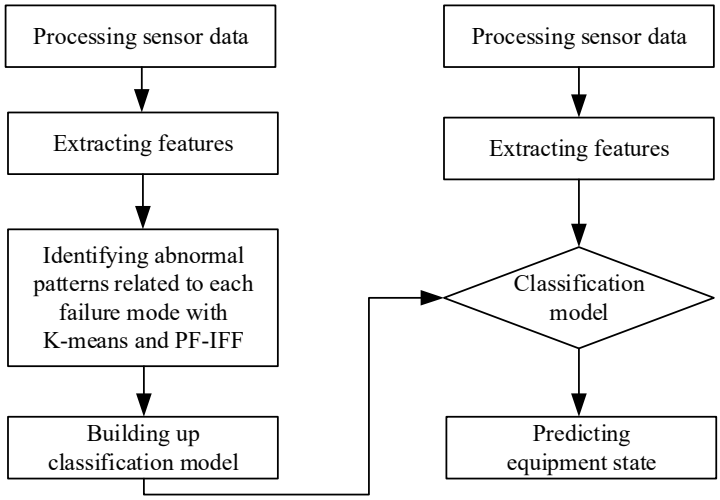

Figure 1. Proposed approach.

3.1. Learning Process

The learning process consisted of four steps. In the first step of the learning process, the data collected from the sensors attached to the equipment were processed by the appropriate technique for learning. In the second step, the feature vector was constructed by dividing the data interval and extracting the summation values for each interval. In the third step, abnormal and normal patterns according to the state of the equipment were extracted using the K-means clustering method and the Pattern Frequency–Inverse Failure mode Frequency (PF–IFF) technique-. A more detailed description of the PF–IFF is given in Section 3.1.3. In the final step, the classification model with the feature vector as input and the state of the equipment as output was built up. With the classification model, the state of the equipment was then able to be diagnosed in the future prediction process.

3.1.1. Processing Data

After collecting the sensor data, at first, it is necessary to refine them for further analysis. To do this, specifically, data clearing or data processing is required. For this, there are several techniques. In this study, we removed the data with missing values and normalized the data. Data normalization means that different scales in each attribute of the data are converted to the same scale. After data normalization, the accuracy and efficiency of the model are known to increase when data mining techniques such as K-means clustering and SVM are applied [25].

3.1.2. Extracting Feature

A feature is an attribute that can be measured individually in an observed phenomenon. A set of these features is called a feature vector. The feature in this study was the summation value of each sensor variable (e.g., temperature, pressure, etc.) in a particular time period. The feature vector is also defined as a set of data summaries for multiple variables during a specific time period. In this study, the summary values of data were set as the mean, standard deviation, maximum, minimum, and median of sensor data in a certain time period. Figure 2 depicts the structure of the feature vector. From each variable, we extracted the summary values of the data at specific time intervals and regarded it as a feature vector.

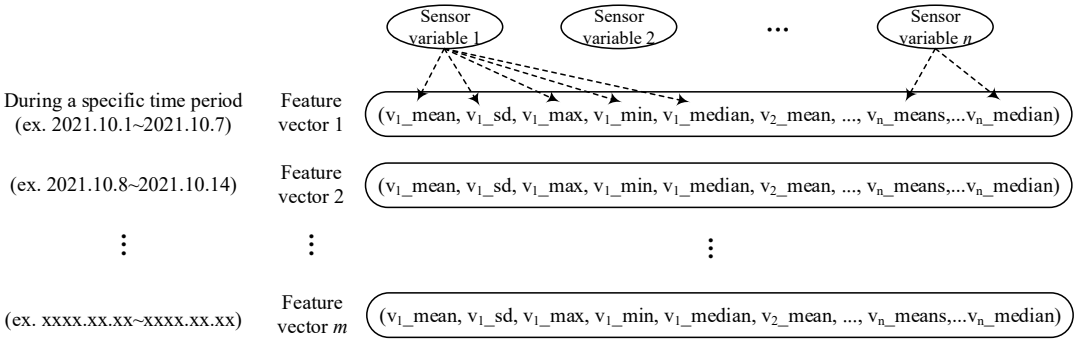

Figure 2. Structure of feature vectors.

3.1.3. Identifying Abnormal Patterns Related to Each Failure Mode

This step was to find the patterns associated with a certain failure mode. In this study, the pattern refers to the cluster of feature vectors. Figure 3 shows the process of identifying abnormal patterns which refer to patterns related to certain failure modes.

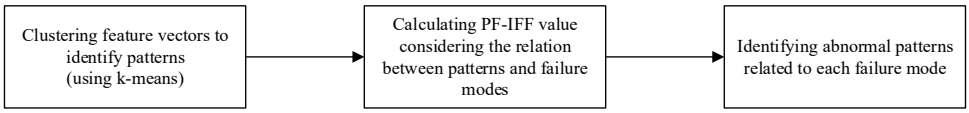

Figure 3. Process for identifying abnormal patterns.

Figure 4 shows the relations among feature vectors, patterns, and failure modes. Briefly, patterns were identified by clustering feature vectors with the K-means algorithm. Then, patterns were associated with a particular failure mode by the PF–IFF method. Detailed descriptions are as follows.

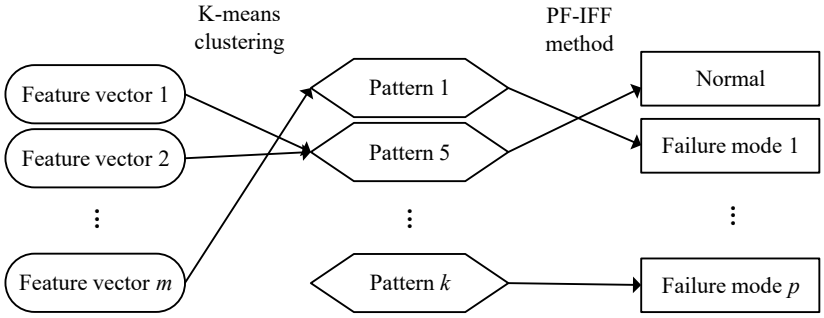

Figure 4. Relations among feature vectors, patterns, and failure modes.

Clustering Feature Vectors to Identify Patterns

In this step, we applied the K-means method [26] to identify a pattern. The K-means divided the data into k clusters, which is a clustering method to minimize the square sum of the distances between the center points of the clusters and the points in the cluster. At this time, the number of clusters (k) suited to the data was obtained using the silhouette technique [27]. The silhouette value had a value between -1 and 1. The larger the value, the better the cluster. Equations (4) and (5) explain how to calculate the silhouette value in a specific feature vector and the mean silhouette value in a specific cluster number:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4}$$

$$\bar{s}(k) = \frac{\sum_{i=1}^{m} s(i)}{m} \tag{5}$$

In this study, it was assumed that the number of initial clusters was larger than the number of states of equipment, and the maximum cluster allowance (r) was determined by a number smaller than the number of feature vectors. The K-means algorithm and the silhouette method were used as follows:

Step 1. Determine the initial value of k in K-means;

Step 2. k feature vectors are arbitrarily extracted from the feature vector set FV = {Feature Vector1, Feature Vector2, . . . , Feature Vector m}, and the extracted feature vectors are set as the centers of clusters;

Step 3. For each feature vector in FV, the distance to the center of k clusters is obtained, and each feature vector is assigned to the cluster having the smallest distance from the center of each cluster;

Step 4. Update the center of the cluster generated in Step 2 by calculating the averages of features included in feature vectors in the same cluster;

Step 5. Steps 3 and 4 are repeated until the grouping of each feature vector is not changed;

Step 6. Calculate the mean silhouette value when the number of clusters is k and repeat steps 2 to 6 until k becomes r with $k = k + 1$;

Step 7. The result of clustering is derived when the average silhouette value at k is maximum.

Calculating the Relations between Patterns and Failure Modes with PF–IFF

In this step, the PF–IFF value was calculated to identify which pattern was more closely associated with a certain failure mode. The PF–IFF value was made based on the TF–IDF value. The TF–IDF is a statistical value indicating how important a particular word is in a particular document when there is a document set consisting of a plurality of documents. In this study, we used the PF–IFF value as a measure of how important a pattern is in a given failure mode in the presence of several failure modes. The PF–IFF value was calculated by Equations (6)–(8) as follows.

$$pf(p,f) = 0.5 + \frac{0.5 \times f(p,f)}{\max\{f(c,f) : c \in f\}} \tag{6}$$

$$iff(p,F) = \log(1 + \frac{|F|}{|\{f \in F : p \in f\}|}) \tag{7}$$

$$pfiff(p,f,F) = pf(p,f) \times iff(p,F) \tag{8}$$

Here, $pf(p,f)$ means the degree of occurrence of a specific pattern p in a certain failure mode f, and $iff(p,F)$ indicates the degree of the common occurrence of a specific pattern p in a failure mode set F. Here, the reason why 1 is included in the log function in (7) is that when a specific pattern appears in all failure modes, the denominator ratio becomes 1 and log(1) = 0, and when multiplied by the pf value, there is no meaning, so 1 is added. $pfiff(p,f,F)$ is the PF–IFF value telling us how important the pattern is in a certain failure mode in the presence of several failure modes.

Identifying Abnormal Patterns Related to Each Failure Mode

In this step, with the PF–IFF calculations for all patterns and failure modes, we determined how many patterns in each failure mode were related to the failure mode. When the number of associated anomalous patterns was set for each failure mode, the patterns having a high PF–IFF value in a certain failure mode were regarded as failure patterns related to the corresponding failure modes.

3.1.4. Learning Classification Model

As shown in Figure 5, in this step, the relationships between the feature vectors and the matching equipment states were learned using a machine learning classifier, for example, SVM, neural network, logistic regression classifier, and so on. In this study, the classifier

was learned based on the states of the equipment (as an output) assigned to the feature vectors (as an input).

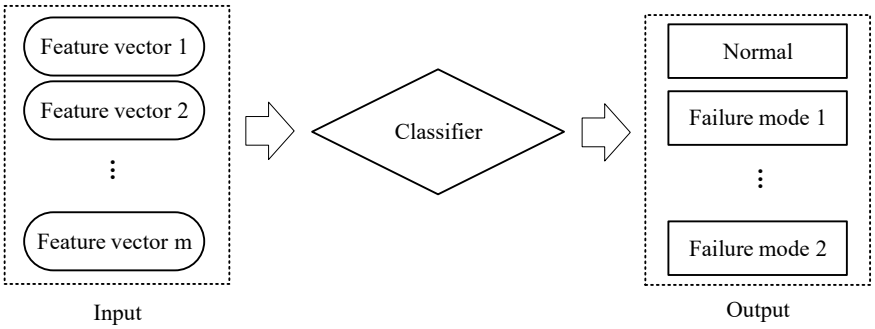

Figure 5. Learning model for classifying the state of equipment.

3.2. Prediction Process

In this step, we diagnosed the condition of equipment and predicted the equipment state using newly collected data. As in the learning process, we processed the newly collected data and extracted feature vectors from it. Then, with the learned classification model, we predicted the state of the equipment.

4. Case Study

In this study, we carried out a case study based on the sensor data gathered from a centrifugal pump of a Korean company *A* in order to prove the validity and effectiveness of the proposed approach. The centrifugal pump is a basic device widely used throughout the industry, such as the equipment industry and the plant industry. It has an energy conversion function that transfers mechanical energy to the working fluid. The failure modes considered in this study were leakage, misalignment, and bearing damage. Leaking means the leakage of gas or liquid in piping or containers. Most of the failures in the mechanical seals of the centrifugal pumps were due to leakage because of primary seal wear [28]. Misalignment means that the mechanical parts are not aligned well with each other. Investigating the defects of rotating equipment, including the centrifugal pump, has shown that misalignment is one main cause of defects in bearings, shafts, impellers, and couplings for rotating bearings [29]. Bearing is a mechanical element that rotates the shaft of the rotating machine while holding the shaft in a fixed position and rotates the shaft while supporting the load on the shaft. Bearing damage generates heat and vibrates the centrifugal pump causing abnormalities [30].

The case study was conducted based on sensor data gathered for four months from June 2013 at intervals of 10 min. The gathered data included the vibration, temperature, speed, and axial movement attributes of each of the 15 pump positions. To carry out the computational experiments accurately, we randomly generated more data based on actual data from June 2013 to July 2015. As a result, a total of three to four failure modes of bearing damage, leakage, and axial movement occurred in the data for two years. In the case study experiment, these data were randomly divided into a 3:1 ratio and applied to training and testing.

In this study, *R 3.4.0* was used to implement the proposed approach. The K-means clustering method was used in the stats library, and the SVM method was used in the *e1071* library. We separated the collected data into training data and test data with a 3:1 ratio for building a learning model and evaluating the learned model, respectively.

4.1. Processing Data

This step was to process data and select variables to improve the performance of the K-means clustering and SVM. The gathered sensor data had several variables on the vibration, temperature, speed, and axial movement for each centrifugal pump. There were a total of 15 variables: 6 vibration, 5 temperature, 2 speed, and 2 axial motion variables.

Figure 6 depicts the data of 4 out of 15 variables. In this study, the tachometer and overspeed variables were excluded from the variable selection because they showed little or no variation except for some abnormalities throughout the data. Therefore, this study considered a total of 13 variables for vibration, temperature, and axial movement. Next, data were processed using the standard z-score to normalize the data.

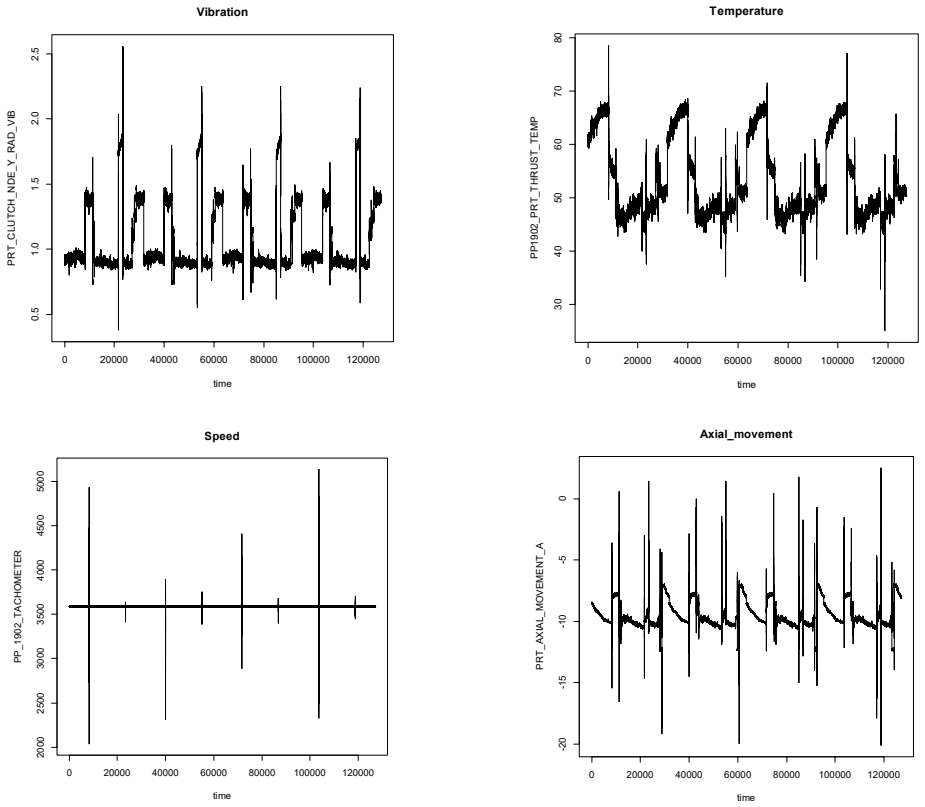

Figure 6. Visualization of gathered sensor data.

4.2. Extracting Feature

In our case study, feature vectors were constructed with summary values (mean, standard deviation, maximum, minimum, median) of each variable every 6 h. Figure 7 shows an example extracting summary values for each variable and generating a set of features as a feature vector.

Vib1	Vib2	Vib3	Vib4	Vib5	Vib6	AXIAL MOVEMENT1	AXIAL MOVEMENT2	TEMP1	TEMP2	TEMP3	TEMP4	TEMP5
0.934	0.872	2.289	2.259	1.075	1.201	–8.539	–8.409	60.635	39.083	55.680	54.018	38.357
0.941	0.878	2.273	2.262	1.078	1.197	–8.536	–8.424	60.677	39.088	55.683	53.975	38.081
0.940	0.878	2.285	2.271	1.072	1.207	–8.535	–8.416	60.564	39.183	55.626	53.952	37.906
…	…	…	…	…	…	…	…	…	…	…	…	…
0.935	0.876	2.296	2.268	1.079	1.202	–8.540	–8.418	60.844	39.101	55.657	54.065	37.859

Vib1_mean	Vib1_sd	Vib1_max	Vib1_min	Vib1_median	Vib2_mean	Vib2_sd	Vib2_max	Vib2_min	Vib2_median	…
–0.510	–0.269	–0.544	–0.464	–0.508	–0.605	–0.485	–0.707	–0.501	–0.603	…
–0.506	–0.104	–0.505	–0.490	–0.505	–0.593	–0.236	–0.636	–0.533	–0.581	…
–0.592	0.487	–0.518	–0.656	–0.559	–0.614	–0.223	–0.636	–0.590	–0.615	…
…	…	…	…	…	…	…	…	…	…	…
–0.575	–0.168	–0.606	–0.537	–0.573	–0.637	–0.421	–0.695	–0.525	–0.636	…

Figure 7. Feature extraction.

4.3. Identifying Abnormal Patterns Related to Each Failure Mode

As mentioned in Section 3.1.3, in this step, we found patterns associated with each failure mode. To this end, first, we made the groups of feature vectors into patterns and calculated the PF–IFF values to identify the patterns associated with the failure modes. Details are as follows:

4.3.1. Clustering Feature Vectors to Identify Patterns

In this step, a pattern was generated by the K-means. For the K-means, deciding the suitable number k of clusters of the feature vectors (patterns) is needed. For this purpose, we used the silhouette method.

Figure 8 shows the average silhouette value for each pattern. Theoretically, the silhouette value has a value between −1 and 1, and the closer to 1, the better the cluster. Since the average silhouette value was the highest at $k = 17$, the number of patterns was determined to be 17 in this study. Figure 9 shows clustering results for vibration, temperature, and axial movement data. The color of the points represented in the figure is as follows. Blue means normal, black means associated with bearing damage, red means associated with leakage, and green means associated with misalignment.

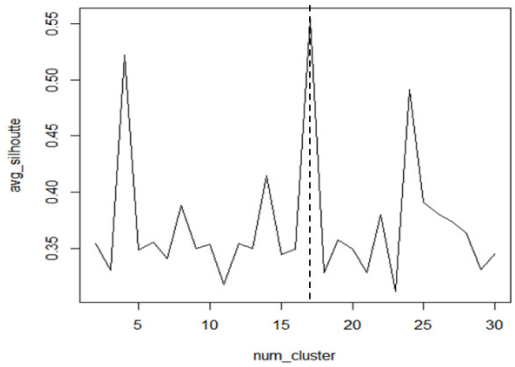

Figure 8. Average silhouette value for each pattern.

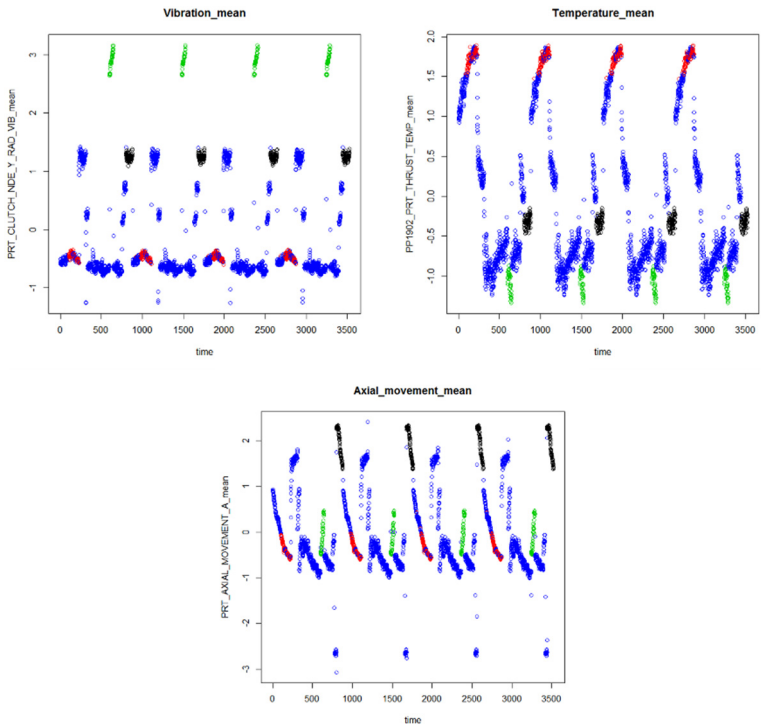

Figure 9. Clustering results for each pattern.

4.3.2. Calculating the Relationship between Patterns and Failure Modes

In this step, the PF–IFF value, which is a criterion for identifying the pattern associated with the failure mode, was calculated. A high PF–IFF value means that a particular pattern is highly relevant for a particular failure mode. Table 4 shows the results of calculating the PF–IFF values between the failure modes (bearing damage, leakage, misalignment) and the cluster of feature vectors in this study.

Table 4. PF–IFF results.

Pattern	Bearing Damage	Leakage	Misalignment
1	0.532 *	0.458	0.793
2	0.494	0.458	0.54
3	0.693	**1.232**	0.693
4	0.49	0.458	0.546
5	0.464	0.458	**0.916**
6	0.567	0.458	0.848
7	0.365	0.362	0.358
8	0.458	0.542	0.472
9	**0.883**	0.693	0.693
10	0.693	**1.297**	0.693
11	0.693	0.904	0.693
12	**1.386**	0.693	0.693
13	0.462	**0.916**	0.458
14	0.693	0.693	**1.238**
15	0.695	0.458	**0.901**
16	0.561	0.458	0.752
17	**0.881**	0.693	0.693

*: PF–IFF value.

4.3.3. Identifying Abnormal Patterns Related to Each Failure Mode

This step identified the associated anomalous pattern based on the PF–IFF value for each failure mode. PF–IFF has a high value when the pattern frequently occurs at a specific failure mode. The bold in Table 4 indicates three patterns having high values associated with each failure mode, for example, pattern 12, pattern 9, and pattern 17 for bearing damage; pattern 10, pattern 3, and pattern 13 for leakage; pattern 14, pattern 5, and pattern 15 for misalignment. Except for the anomalous patterns selected, as above, the others are considered normal patterns.

In this study, we assumed that the number of normal patterns is larger than the number of abnormal patterns in all collected data and limited the maximum number of abnormal patterns in each failure mode to 3. Therefore, the classifier was trained for a total of 27 cases by varying the number of abnormal patterns in each failure type, and its performance was checked.

Figure 10 depicts several variables to see how feature vectors associated with failure modes are distributed when there is one associated pattern in each failure mode. Misalignment (green color) seems to have a high value in vibration. Leakage (red color) seems to have a relatively high value at temperature and bearing damage (black color) seems to have a high value in axial movement.

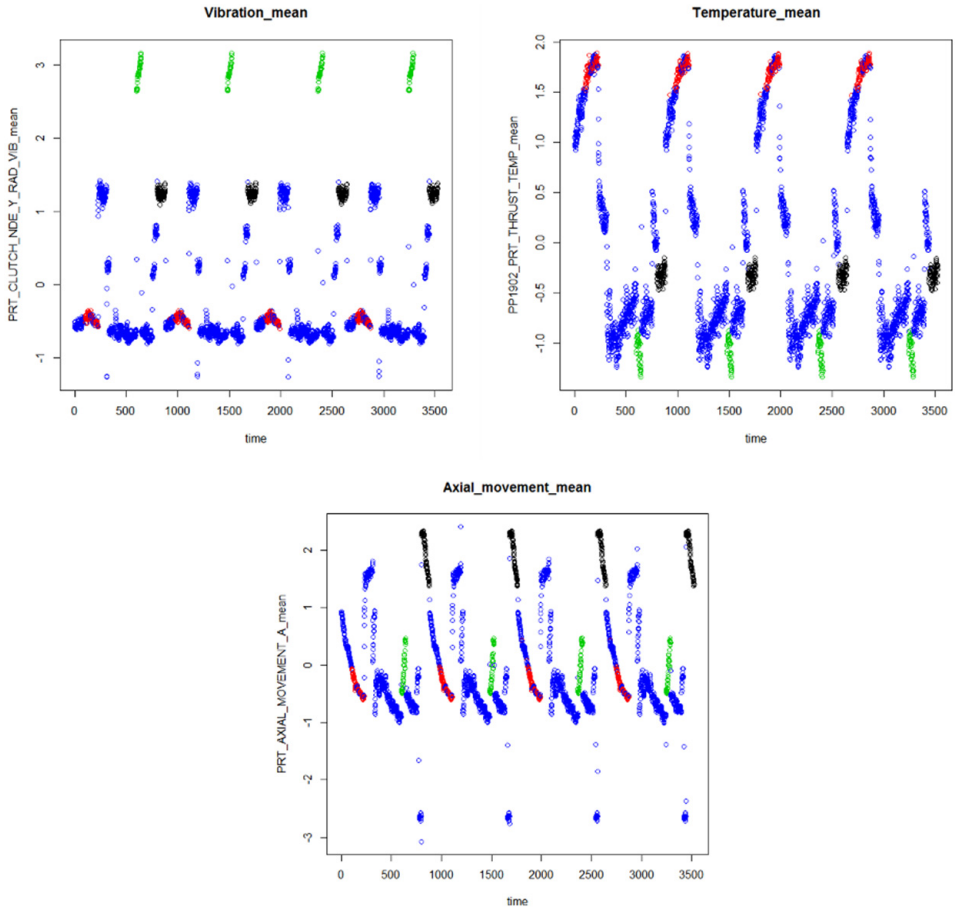

Figure 10. Scatter diagrams of equipment state.

4.4. Learning Classification Model

In the case study, SVM [31] was used as the classifier. The SVM is a technique that learns the classifier by maximizing the margin of the hyperplane and the support vector. The SVM was learned from training data by using the feature vector and the state of the equipment as inputs and outputs, respectively. Figure 11 is a schematic diagram of how the classification model was learned.

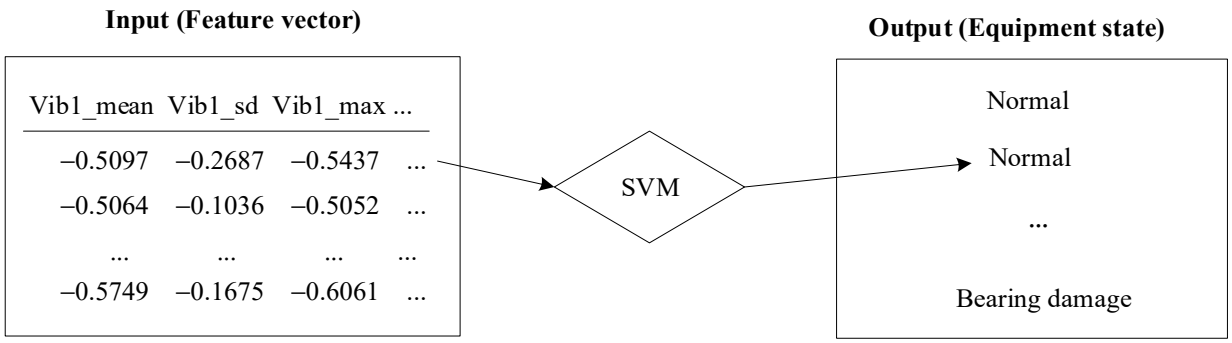

Figure 11. Classification model learning through SVM classifier.

4.5. Evaluation

The performance of the proposed PF–IFF-based approach was evaluated by comparing the performance with and without PF–IFF when the SVM classifier was applied to classify the failure modes. In this study, the actual state of the equipment was given in order to evaluate the performance of the SVM. The state of the actual equipment corresponding to each feature vector was set to the state of the equipment failure mode from the point

of time when the history was recorded in the specific type of equipment to the specific point of time. Figure 12 depicts scatter plots showing the actual state of the equipment for 3 of the 13 variables in the test data. The PF–IFF-based method was used to find anomalies associated with the type of equipment failure and to assign the state of the equipment to each feature vector. In Figure 12, the red color represents the state for leakage, green represents misalignment, black represents bearing damage, and blue represents the steady state. The vertical line shown in Figure 12 indicates the time when each type of failure occurred.

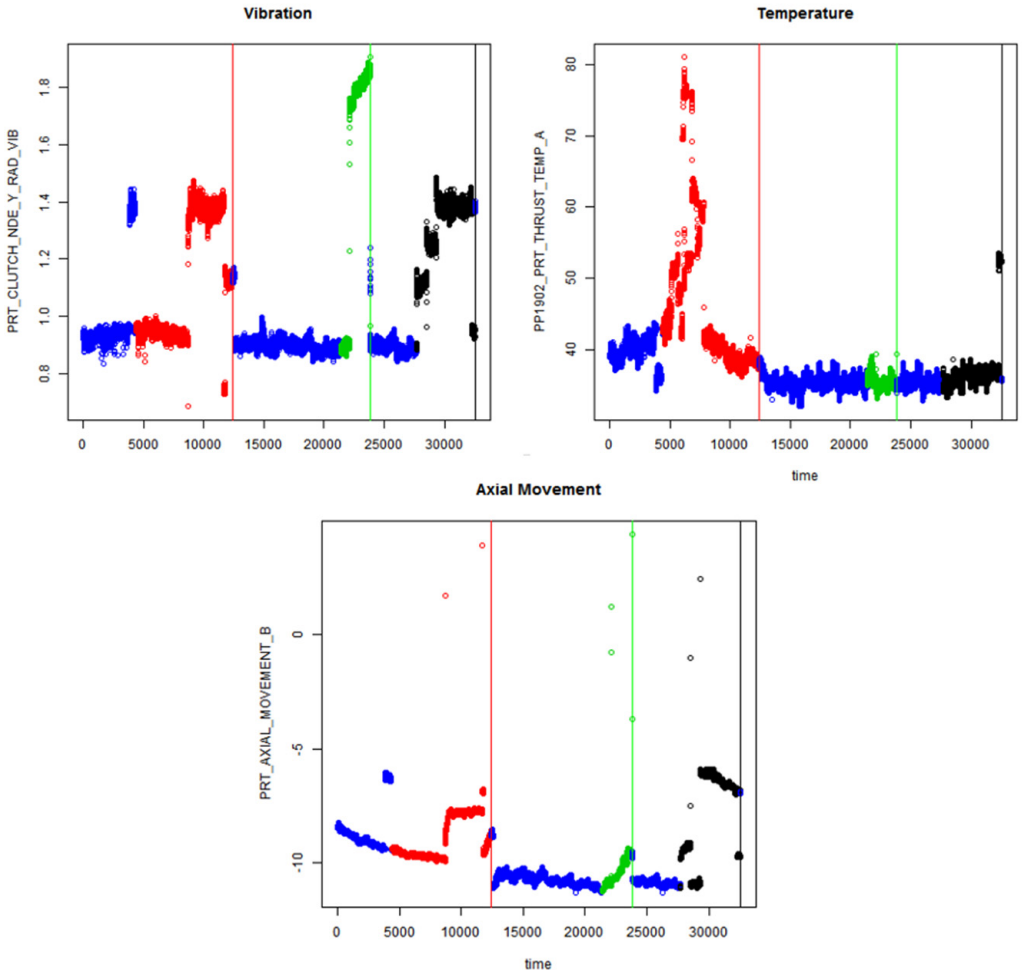

Figure 12. Scatter plots of actual states of the equipment.

However, when the PF–IFF is not used, it is necessary to allocate the state of the equipment to each feature vector when learning the classifier. To this end, we assigned the state of the equipment to each feature vector with the failure mode or normal identifier using the equipment status data during a certain period before the occurrence of the specific failure mode. For certain periods, we used 1, 2, 3, and 4 days for the computational tests, respectively, in order to know which day was more suitable for classification. In the experiment using PF–IFF, performance was evaluated for the cases where the number of abnormal patterns associated with the failure mode was different, as mentioned in Section 4.3.3.

The performance criterion was set as Area Under the Curve (AUC) in the Receiver Operating Characteristic (ROC) Curve. ROC Curve is a technique for evaluating the performance of classification using sensitivity and specificity. For example, sensitivity refers to the probability of being correctly classified as leakage when there is actual leakage, and specificity means the probability that the leakage is correctly classified as not leakage. The ROC curve and the AUC were calculated using the above two values for each state of

the equipment to evaluate the performance of the model. In the ROC curve, the greater the degree of sensitivity increase, the larger the area of AUC. The closer the AUC is to 1, the better the performance is. The values in Tables 5–7 are these AUC values.

Table 5. Classified results of SVM when PF–IFF is not used.

Criterion for State Assignment	Normal	Bearing Damage	Leakage	Misalignment
Day 1	**0.722 ***	**0.593**	**0.739**	**0.689**
Day 2	0.544	0.521	0.563	0.398
Day 3	0.500	0.466	0.548	0.499
Day 4	0.573	0.511	0.744	0.615

*: AUC value.

Table 6. Classified results when PF–IFF was used.

Combination of the Number of Patterns for Each Failure Mode	Normal	Bearing Damage	Leakage	Misalignment
BearingDamage:1, Leakage:1, Misalignment:1	0.788 *	0.895	0.544	0.793
BearingDamage:1, Leakage:1, Misalignment:2	0.679	0.86	0.541	0.813
BearingDamage:1, Leakage:1, Misalignment:3	0.624	0.863	0.477	0.892
BearingDamage:1, Leakage:2, Misalignment:1	0.932	0.880	0.914	0.758
BearingDamage:1, Leakage:2, Misalignment:2	0.837	0.879	0.901	0.789
BearingDamage:1, Leakage:2, Misalignment:3	0.791	0.892	0.879	0.889
BearingDamage:1, Leakage:3, Misalignment:1	0.744	0.920	0.806	0.737
BearingDamage:1, Leakage:3, Misalignment:2	0.656	0.912	0.791	0.771
BearingDamage:1, Leakage:3, Misalignment:3	0.605	0.946	0.779	0.887
BearingDamage:2, Leakage:1, Misalignment:1	0.783	0.944	0.542	0.791
BearingDamage:2, Leakage:1, Misalignment:2	0.699	0.929	0.537	0.807
BearingDamage:2, Leakage:1, Misalignment:3	0.659	0.943	0.482	0.891
BearingDamage:2, Leakage:2, Misalignment:1	0.950	0.940	0.909	0.757
BearingDamage:2, Leakage:2, Misalignment:2	0.861	0.929	0.899	0.785
BearingDamage:2, Leakage:2, Misalignment:3	0.827	0.968	0.887	0.889
BearingDamage:2, Leakage:3, Misalignment:1	0.754	0.949	0.802	0.731
BearingDamage:2, Leakage:3, Misalignment:2	0.678	0.946	0.791	0.768
BearingDamage:2, Leakage:3, Misalignment:3	0.635	0.979	0.783	0.889
BearingDamage:3, Leakage:1, Misalignment:1	0.832	0.976	0.584	0.798
BearingDamage:3, Leakage:1, Misalignment:2	0.755	0.979	0.547	0.805
BearingDamage:3, Leakage:1, Misalignment:3	0.714	0.982	0.463	0.891
BearingDamage:3, Leakage:2, Misalignment:1	**0.957**	**0.977**	**0.930**	**0.765**
BearingDamage:3, Leakage:2, Misalignment:2	0.888	0.976	0.909	0.787
BearingDamage:3, Leakage:2, Misalignment:3	0.860	0.981	0.891	0.888
BearingDamage:3, Leakage:3, Misalignment:1	0.754	0.963	0.813	0.750
BearingDamage:3, Leakage:3, Misalignment:2	0.695	0.969	0.804	0.774
BearingDamage:3, Leakage:3, Misalignment:3	0.661	0.974	0.794	0.891

*: AUC value.

Table 7. Comparison results.

Whether PF–IFF Was Used or Not	Normal	Bearing Damage	Leakage	Misalignment
Used	**0.957** *	**0.977**	**0.930**	**0.765**
Not used	0.541	0.916	0.818	0.824

*: AUC value.

Table 5 shows the experimental results using SVM when PF–IFF was not used. As the test results show, when the failure mode was assigned with the equipment status data one day before the occurrence of the specific failure mode, the result is the best. It shows that AUC was 0.722 (normal), 0.593 (bearing damage), 0.739 (leakage), and 0.689 (misalignment), respectively.

Table 6 shows the AUC of each state of the equipment when the number of abnormal patterns associated with the failure mode was changed. We endeavored to know whether the optimum performance is achieved at a certain setting by comparing the AUC and the comparison result when the PF–IFF was not used. The number of abnormal patterns associated with each type of failure was increased up to 3, and the experiment was conducted 27 times to evaluate the performance of the PF–IFF-based method with the SVM classifier. For example, Table 7 shows that the AUC for each state of the equipment (normal, bearing damage, leakage, misalignment) was 0.957, 0.977, 0.930, and 0.765 when there were three, two, and one abnormal pattern/s corresponding to bearing damage, leakage, and misalignment, which is the best performance case. In this case, the abnormal patterns in each failure mode were pattern 9, pattern 12, and pattern 17 in the case of bearing damage, pattern 3 and pattern 10 in the case of leakage, and pattern 5, pattern 14, and pattern 15 in the case of misalignment.

Table 7 compares the performance of the model according to the use of PF–IFF based on AUC. The best performance for identifying the state of the equipment was 0.957 (normal), 0.977 (bearing damage), 0.930 (leakage), and 0.765 (misalignment), when the PF–IFF was used. Overall, we know that the PF–IFF method gave better performance in identifying the state of equipment than the unused method did.

5. Conclusions

This study has proposed an approach for diagnosing failure modes based on the sensor data gathered from a centrifugal pump. In the proposed approach, the PF–IFF-based method using TF–IDF was proposed to extract the specific patterns for each mode of failure of the equipment. Furthermore, the SVM with the PF–IFF method was applied to identify the pattern of the failure mode. The case study on the centrifugal pump has been conducted to identify the failure modes such as 'Bearing Damage', 'Leakage', and 'Misalignment' based on the sensor data associated with vibration, temperature, and axial movement. Experimental results showed that the PF–IFF-based SVM approach gave better performance in diagnosing the equipment failures than the SVM approach.

Although the proposed approach has meaningful results, there are several limitations, as follows. First, the proposed approach could tell us only what kind of failure will occur when an anomaly pattern is found in the equipment. It cannot predict the remaining lifespan, i.e., when the actual failure will occur. Second, the approach proposed in this study focuses on the diagnosis and prediction of failures of relatively simple systems, and it is considered that it is difficult to apply to more complex systems or to complex situations in which the phenomena or types of failures are not simply analyzed. Third, it should also be pointed out that the high AUC values in the experimental results of this study may be a result of overfitting the learning model. This is believed to be due to the phenomenon caused by the small amount of data used in the experiment of this study and having an imbalance. In the future, based on more real data, we will need to tune the training model to avoid overfitting. Finally, the last limitation is that certain patterns that appear in the

interval can be diagnosed as normal if they occur in all types of failures in the equipment, even though there is an abnormality in the condition of the equipment at a particular time period. In future studies, the following could be considered. For example, it is necessary to research methods for estimating failure times according to equipment failure modes and for detecting common patterns that occur before equipment failure when diagnosing equipment conditions.

Author Contributions: Conceptualization, M.-Y.S. and H.-B.J.; Data curation, S.-Y.H., J.-H.L. and H.-B.J.; Formal analysis, M.-Y.S., S.-Y.H. and H.-B.J.; Funding acquisition, H.-B.J.; Investigation, M.-Y.S., S.-Y.H., J.-G.K. and H.-B.J.; Methodology, M.-Y.S. and H.-B.J.; Project administration, H.-B.J.; Supervision, H.-B.J.; Validation, M.-Y.S. and H.-B.J.; Writing—original draft, S.-Y.H., J.-H.L., J.-G.K. and H.-B.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1F1A1063058).

Data Availability Statement: The data presented in this study are not publicly available due to the company's security issue.

Conflicts of Interest: The authors declare no conflict of interest.

Notation

i	index for feature vector
p	index for each pattern
f	index for each failure mode
F	a set of failure modes
k	the number of clusters (patterns)
n	the number of variables
m	the number of feature vectors
FV	a set of feature vectors
r	the number of maximum allowable clusters (patterns)
$a(i)$	average distance between the feature vector i and other points in the same cluster
$b(i)$	average distance of the feature vector i for the points in the other clusters
$s(i)$	silhouette value in the feature vector i
$\bar{s}(k)$	average silhouette value when the number of clusters is k
$f(p, f)$	the total frequency of pattern p in failure mode f
$\max\{f(c, f) : c \in f\}$	the frequency of the largest pattern in a certain failure mode
$\|F\|$	size of failure mode set F (total number of failure modes)
$\|\{f \in F : p \in f\}\|$	the number of failure modes, including the particular pattern p.

References

1. Arunraj, N.S.; Maiti, J. Risk-based maintenance policy selection using AHP and goal programming. *Saf. Sci.* **2010**, *48*, 238–247. [CrossRef]
2. Shin, J.H.; Jun, H.B. On condition based maintenance policy. *J. Comput. Des. Eng.* **2015**, *2*, 119–127. [CrossRef]
3. Bunks, C.; McCarthy, D.; Al-Ani, T. Condition-based maintenance of machines using hidden Markov models. *Mech. Syst. Signal Process.* **2000**, *14*, 597–612. [CrossRef]
4. Djurdjanovic, D.; Lee, J.; Ni, J. Watchdog Agent—An infotronics-based prognostics approach for product performance degradation assessment and prediction. *Adv. Eng. Inform.* **2003**, *17*, 109–125. [CrossRef]
5. Jardine, A.K.S.; Lin, D.; Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **2006**, *20*, 1483–1510. [CrossRef]
6. Salton, G.; Fox, E.A.; Wu, H. Extended Boolean information retrieval. *Commun. ACM* **1983**, *26*, 1022–1036. [CrossRef]
7. Phelan, O.; McCarthy, K.; Smyth, B. Using twitter to recommend real-time topical news. In Proceedings of the Third ACM Conference on Recommender Systems, New York, NY, USA, 23–25 October 2009; pp. 385–388.
8. Go, G.S.; Jung, W.K.; Shin, Y.G.; Park, S.S.; Jang, D.S. A Study on Development of Patent Information Retrieval Using Text mining. *J. Korea Acad.-Ind. Coop. Soc.* **2011**, *12*, 3677–3688.
9. Trstenjak, B.; Mikac, S.; Donko, D. KNN with TF-IDF based Framework for Text Categorization. *Procedia Eng.* **2014**, *69*, 1356–1364. [CrossRef]

10. Yoo, E.S.; Choi, G.H.; Kim, S.H. Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels. *J. Korea Soc. Comput. Inf.* **2015**, *20*, 121–129. [CrossRef]
11. Dong, M.; He, D. Hidden semi-Markov model-based methodology for multi-sensor equipment health diagnosis and prognosis. *Eur. J. Oper. Res.* **2007**, *178*, 858–878. [CrossRef]
12. Wang, H.Q.; Chen, P. Fault diagnosis of centrifugal pump using symptom parameters in frequency domain. *Agric. Eng. Int. CIGR J.* **2007**, *9*, 1–14.
13. Yang, M.D.; Su, T.C. Automated diagnosis of sewer pipe defects based on machine learning approaches. *Expert Syst. Appl.* **2008**, *35*, 1327–1337. [CrossRef]
14. Fei, S.W.; Zhang, X.B. Fault diagnosis of power transformer based on support vector machine with genetic algorithm. *Expert Syst. Appl.* **2009**, *36*, 11352–11357. [CrossRef]
15. Li, Y.G.; Nilkitsaranont, P. Gas turbine performance prognostic for condition-based maintenance. *Appl. Energy* **2009**, *86*, 2152–2161. [CrossRef]
16. Son, J.H.; Ko, J.M.; Kim, C.O. Feature Based Decision Tree Model for Fault Detection and Classification of Semiconductor Process. *IE Interfaces* **2009**, *22*, 126–134.
17. Zhou, R.; Bao, W.; Li, N.; Huang, X.; Yu, D. Mechanical equipment fault diagnosis based on redundant second generation wavelet packet transform. *Digit. Signal Process.* **2010**, *20*, 276–288. [CrossRef]
18. Wu, B.; Tian, Z.; Chen, M. Condition-based maintenance optimization using neural network-based health condition prediction. *Qual. Reliab. Eng. Int.* **2013**, *29*, 1151–1163. [CrossRef]
19. Muralidharan, V.; Sugumaran, V.; Indira, V. Fault diagnosis of monoblock centrifugal pump using SVM. *Eng. Sci. Technol. Int. J.* **2014**, *17*, 152–157. [CrossRef]
20. Moon, D.S.; Kim, S.H. Development of intelligent fault diagnostic system for mechanical element of wind power generator. *J. Korean Inst. Intell. Syst.* **2014**, *24*, 78–83.
21. Kim, Y.G.; Cho, S.J.; Jun, H.B.; Ha, J.H.; Shin, J.H. A Study on Fault Prediction Method in a Pump Tower of LNG FPSO. *Korean J. Comput. Des. Eng.* **2016**, *21*, 111–121. [CrossRef]
22. Mostafa, S.A.; Mustapha, A.; Hazeem, A.A.; Khaleefah, S.H.; Mohammed, M.A. An agent-based inference engine for efficient and reliable automated car failure diagnosis assistance. *IEEE Access* **2018**, *6*, 8322–8331. [CrossRef]
23. Contreras-Valdes, A.; Amezquita-Sanchez, J.P.; Granados-Lieberman, D.; Valtierra-Rodriguez, M. Predictive data mining techniques for fault diagnosis of electric equipment: A review. *Appl. Sci.* **2020**, *10*, 950. [CrossRef]
24. Abid, A.; Khan, M.T.; Iqbal, J. A review on fault detection and diagnosis techniques: Basics and beyond. *Artif. Intell. Rev.* **2021**, *54*, 3639–3664. [CrossRef]
25. Ng, M.K.; Li, M.J.; Huang, J.Z.; He, Z. On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 503–507. [CrossRef]
26. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc.* **1979**, *28*, 100–108. [CrossRef]
27. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
28. Jeoung, R.H.; Lee, B.K. Fault Detection Signal for Mechanical Seal of Centrifugal Pump. *J. Korean Soc. Saf.* **2012**, *27*, 20–27.
29. Jeoung, R.H.; Chai, J.B.; Lee, B.H.; Lee, D.H.; Lee, B.K. Feature parameter analysis for rotor fault diagnosis. *KSFM J. Fluid Mach.* **2012**, *15*, 31–38. [CrossRef]
30. Tinga, T. *Mechanism Based Failure Analysis*; Netherlands Defense Academy Publisher: Den Helder, The Netherlands, 2012.
31. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]