Research Article

Optimal Statistical Feature Subset Selection for Bearing Fault Detection and Severity Estimation

Chhaya Grover ^{1,2} and Neelam Turk¹

¹Department of Electronics Engineering, J.C. Bose University of Science and Technology (YMCA), Sector-6, Faridabad, Haryana 121006, India

²Department of Electronics and Communication Engineering, JSS Academy of Technical Education, Sector-62, Noida, U.P. 201301, India

Correspondence should be addressed to Chhaya Grover; chhayagrover@jssaten.ac.in

Received 4 February 2020; Revised 23 July 2020; Accepted 1 August 2020; Published 26 August 2020

Academic Editor: Anil Kumar

Copyright © 2020 Chhaya Grover and Neelam Turk. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The performance of bearing fault detection systems based on machine learning techniques largely depends on the selected features. Hence, selection of an ideal number of dominant features from a comprehensive list of features is needed to decrease the number of computations involved in fault detection. In this paper, we attempted statistical time-domain features, namely, Hjorth parameters (activity, mobility, and complexity) and normal negative log likelihood for Gaussian mixture model (GMM) for the first time in addition to 26 other established statistical features for identification of bearing fault type and severity. Two datasets are derived from a publicly available database of Case Western Reserve University to identify the capability of features in fault identification under various fault sizes and motor loads. Features have been investigated using a two-step approach—filter-based ranking with 3 metrics followed by feature subset selection with 11 search techniques. The results indicate that the set of features root mean square, geometric mean, zero crossing rate, Hjorth parameter—mobility, and normal negative log likelihood for GMM outperforms other features. We also compared the diagnostic performance of normal negative log likelihood for GMM with the established feature normal negative log likelihood for single Gaussian. The selected set of statistical features is validated using ensemble rule-based classifiers and showed an average accuracy of 96.75% with proposed statistical features subset and 99.63% with all 30 features. *F*-measure and *G*-mean scores are also calculated to investigate their performance on datasets with class imbalance. The diagnostic effectiveness of the features was further validated on a bearing dataset obtained from an operating thermal power plant. The results obtained show that our newly proposed feature subset plays a major role in achieving good classification results and has a future potential of being used in a high-dimensional dataset with multidomain features.

1. Introduction

Vibration-based diagnosis and prognosis of bearing faults has been an active area of research for many years [1–3]. Machine learning algorithms have been effectively used in bearing fault detection systems. The performance of such algorithms, in terms of accuracy and efficiency, largely depends on the number of features used and their discriminating capability.

Several types of statistical features have been used so far. Zeng et al. [4] used 12 conventional time-domain and 10 frequency-domain features for their classification approach.

The conventional statistical parameters, namely, mean value, maximum value, root mean square value, peak-to-peak value, root, standard deviation, variance, kurtosis, and skewness are the most commonly used dimensional parameters [5–8]. A wide set of dimensionless statistical parameters such as crest factor, shape factor, impulse factor, clearance factor, skewness factor, and kurtosis factor has also been used to enhance the performance of fault diagnosis systems [9–12]. These parameters are extracted from either raw or preprocessed vibration signals. For the latter, empirical mode decomposition (EMD) has been popularly used as a preprocessing technique [13]. It is important to note that

not all the constituent signals, called intrinsic mode functions (IMFs) generated by EMD, bear fault information. Different criteria, such as correlation coefficient [14] and kurtosis [4], have been used to select significant IMFs. This has been shown to be important for designing an efficient fault diagnosis system [15, 16].

A major boost in bearing fault diagnosis research was seen when novel statistical features began to be applied in artificial intelligence techniques. Sreejith et al. [17] proposed a method for pattern recognition by extracting normal negative log likelihood and kurtosis values of time-domain signals and using them as input to an artificial neural network. Abbasion et al. [18] used Weibull negative log-likelihood function of denoised signals and support vector machine (SVM) classifier for fault classification of rolling bearings, while in [19–21] zero crossings, mean absolute deviation, and histogram features were applied in their respective works. The application of a new group of statistical parameters, called as “Hjorth parameters,” was explored for diagnosing 4 types of faults in rolling element ball bearings [22]. Lastly, different kinds of entropies have also been used for intelligent fault diagnosis in various research studies [14, 23–25]. Some recent papers have developed symmetric cross entropy measures of single-valued neurotrophic sets for fault diagnosis in bearings [26, 27].

Researchers were thus able to utilise the power of statistical features for capturing important fault information by applying popular supervised machine learning techniques. The potential of one such machine learning technique—rule-based classifiers—in the domain of bearing fault diagnosis has also been shown to be promising. An ensemble of rule-based classifiers for fault diagnosis of rotating machinery was proposed by Dou et al. [28] to predict potential faults and subsequent breakdown of rotating machinery by using 6 time-domain and 5 frequency-domain features. Application of rule-based classifiers has also been successfully extended for severity detection. Li et al. [29] proposed an association rule mining method based on equal probability discretization to perform a 10-class prediction on select time-domain and frequency-domain features for fault classification and severity prediction.

Although all of the studies cited above have applied various time-domain parameters, whether conventional or novel and dimensional or dimensionless, for bearing fault diagnosis, either exclusively or in addition to other frequency-domain and time-frequency features, none of the studies discusses and provides the reason behind the use of a set of particular statistical features. Majority of the previous studies, till now, have focussed more on optimising fault diagnosis classifiers. Moreover, few have compared established features with novel ones. Since fault diagnostic accuracy depends on the choice of both features and classifier, we believe it is important to emphasise on features too by proposing novel statistical features and/or developing a strategy to conduct an extensive analysis of new and old statistical features, so as to derive the most optimal set of fault-sensitive features.

In view of this, this paper applies statistical time-domain features, namely, Hjorth parameters (activity, mobility, and

complexity) and normal negative log likelihood for Gaussian mixture model (GMM) for the first time for rolling element bearing fault type and severity diagnosis. These statistical features were added to 26 already established statistical features to investigate a comprehensive set of 30 time-domain statistical features. The features were extracted from intrinsic mode functions of bearing vibration signals obtained using EMD. An in-depth two-step model-agnostic approach, based on filter-based ranking with 3 metrics and feature subset selection with 11 search techniques including both conventional search techniques and the newly introduced swarm search techniques, has been developed to select an optimal feature subset for bearing fault type detection and severity estimation. For validation of the performance of the selected features, rule-based classifiers—PART and JRIP—and their bagging and boosting ensembles were applied. The paper also examines the selected parameters and further analyzes the replacement of the high ranked parameter normal negative log likelihood for single Gaussian with normal negative log likelihood for GMM and its effect on the diagnostic performance. Case Western Reserve University (CWRU) Bearing Dataset [30] has been used for the experiments. The performances of the selected subset of features have also been tested on a dataset obtained from an operating power plant for fault type detection. The main contribution of this paper is twofold: (1) proposing Hjorth parameters and normal negative log likelihood for Gaussian mixture model (GMM) for rolling element ball bearing fault detection and severity estimation and (2) devising a strategy to probe a comprehensive set of statistical time-domain features with the objective of identifying optimal feature subset for rolling element ball bearing fault detection and severity estimation.

The paper has been organised as follows: Section 2 describes the statistical features and filter-based feature ranking and subset selection methods. Section 3 explains the steps of our proposed methodology. Section 4 discusses in detail data preparation steps, implementation and results of feature selection, fault type detection, and fault severity detection. This section also discusses the replacement of normal negative log likelihood for single Gaussian with its Gaussian mixture model variant as well as the implementation of our methodology on a dataset obtained from an operating power plant. Finally, Section 5 discusses the drawn conclusions and future scope of our research.

2. Theoretical Background

2.1. Extraction of Statistical Features. Table 1 shows the list of statistical features extracted from the vibration signals. In the table, $T_1, T_2, \dots, T_{30}$ are feature notations, x_i are individual data points of sampled vibrational signal, and N is the length of the vibration signal. The dimensional statistical parameters T_1 to T_{11} , T_{18} , and T_{19} and dimensionless statistical parameters T_{12} to T_{17} are self-explanatory from their formulae. T_{20} and T_{21} are the arithmetic means of the absolute deviations of each x_i from mean and median, respectively. T_{22} is the rate of zero crossing (ZCR) and can be defined as the ratio of the number of zero crossings to the total number of points in the signal. In other words, ZCR is the rate at

TABLE 1: List of statistical time-domain features.

Feature
Notation and formula
Mean T1 = $(1/N) \sum_{i=1}^N x_i$
Root mean square T2 = $\sqrt{(1/N) \sum_{i=1}^N x_i^2}$
Root T3 = $((1/N) \sum_{i=1}^N \sqrt{ x_i })^2$
Maximum value T4 = $\max(x)$
Peak-to-peak T5 = $\max(x) - \min(x)$
Standard deviation T6 = $\sqrt{1/(N-1) \sum_{i=1}^N (x_i - \mu)^2}$
Median T7 = (50(N+1)/100)th observation
25th percentiles T8 = (25(N+1)/100)th observation
75th percentiles T9 = (75(N+1)/100)th observation
Skewness T10 = $(1/\sigma^3)((1/(N-1)) \sum_{i=1}^N (x_i - \mu)^3)$
Kurtosis T11 = $(1/\sigma^4)((1/(N-1)) \sum_{i=1}^N (x_i - \mu)^4)$
Crest factor T12 = $(\max(x)/\sqrt{(1/N) \sum_{i=1}^N x_i^2})$
Shape factor T13 = $(\sqrt{(1/N) \sum_{i=1}^N x_i^2} / (1/N) \sum_{i=1}^N x_i)$
Impulse factor T14 = $(\max(x) / (1/N) \sum_{i=1}^N x_i)$
Clearance factor T15 = $(\max(x) / ((1/N) \sum_{i=1}^N \sqrt{ x_i }))$
Skewness factor T16 = $(1/\sigma^3)((1/(N-1)) \sum_{i=1}^N (x_i - \mu)^3 / (\sqrt{(1/N) \sum_{i=1}^N x_i^2})^3)$
Kurtosis factor T17 = $(1/\sigma^4)((1/(N-1)) \sum_{i=1}^N (x_i - \mu)^4 / (\sqrt{(1/N) \sum_{i=1}^N x_i^2})^4)$
Geometric mean T18 = $(\prod_{i=1}^N x_i)^{1/N}$
Root-sum-of-squares (RSSQ) T19 = $\sqrt{\sum_{i=1}^N x_i ^2}$
Mean absolute deviation T20 = $\text{mean}(x_i - \mu)$
Median absolute deviation T21 = $\text{median}(x_i - \text{median}(x_i))$
Rate of zero crossing T22 = (number of zero crossings/total number of points)
Entropy T23 = $-\sum_{i=1}^N h(x_i) \log_2 h(x_i)$
Histogram upper bound T24 = $\max(x_i) + ((\max(x_i) - \min(x_i/(N-1)))/2)$
Histogram lower bound T25 = $\max(x_i) - ((\max(x_i) - \min(x_i/(N-1)))/2)$
Activity T26 = $(1/(N-1)) \sum_{i=1}^N (x_i - \mu)^2$
Mobility T27 = (standard deviation of $x'(n)$ /standard deviation of $x(n)$)
Complexity T28 = (mobility of $x'(n)$ /mobility of $x(n)$)
Weibull negative log likelihood T29 = $-\sum_{i=1}^N \log f(a, b x_i)$
Normal negative log likelihood for single Gaussian T30 = $-\sum_{i=1}^N \log[(1/\sigma\sqrt{2\pi})\exp(-(x_i - \mu)^2/2\sigma^2)]$

which a signal changes sign from negative to positive or vice versa, thus providing a good understanding of the frequency of the signal and hence very useful for separating signals with different frequencies. T23 or entropy is a measure of randomness of the values of a vibration signal. The two histogram features T24 and T25 are histogram upper and lower bounds, respectively. The next three statistical parameters are the Hjorth parameters. T26 or activity of vibration signal is the same as variance of the signal. T27 or mobility is defined as the ratio of the standard deviation of the first derivative of the vibration signal to the standard deviation of the vibration signal itself. T28 or complexity is defined as the ratio of the mobility of the first derivative of the vibration signal to the mobility of the vibration signal. The next statistic, T29 or Weibull negative log likelihood, is computed as follows:

$$T29 = -\log \prod_{i=1}^N f(a, b | x_i) = -\sum_{i=1}^N \log f(a, b | x_i). \quad (1)$$

Finally, T30 or normal negative log likelihood for single Gaussian can be computed as follows:

$$T30 = -\sum_{i=1}^N \log \left[\frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i - \mu)^2}{2\sigma^2} \right]. \quad (2)$$

2.2. Filter-Based Feature Selection Methods. In this paper, we have used filter-based methods for identifying important features for bearing fault diagnosis. Filter-based techniques work well with any classification model producing accurate results with shorter run times [31]. Moreover, these techniques perform well for the identification of dominant features for both fault type and severity detection [32].

2.2.1. Feature Ranking. Three filter-based ranking metrics have been used to rank the features: chi-squared, information gain, and gain ratio. In chi-squared feature selection, the value of a statistic called chi-square (χ_c^2) is calculated for the feature and response variables. Those features that have

high chi-squared value are then selected. The chi-squared value is calculated as follows:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \quad (3)$$

where c = degrees of freedom, O_i = observed value, and E_i = expected value.

Information gain test uses the value of a statistic called information gain that is computed for each feature (L) in the context of a response variable (R), ultimately selecting features with higher information gain values. The higher the information gain score for an attribute, the more is the information that the attribute can contribute to the response variable. Information gain can be understood as the reduction in entropy (H) and is calculated as follows:

$$\text{information gain} = \Delta H = H - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R, \quad (4)$$

where

$$\text{entropy}(H) = - \sum_{k=1}^K p_k \log_2 p_k, \quad (5)$$

and m = the total number of instances and m_k = the number of instances belonging to class k , in which $k = 1, \dots, K$.

The third metric of gain ratio feature is a ratio of information gain to the intrinsic information:

$$\text{gain ratio}(S, A) = \frac{\text{gain}(S, A)}{\text{split entropy}(S, A)}, \quad (6)$$

where

$$\text{split entropy}(S, A) = - \sum_{V \in \text{values}(A)} \frac{|S_V|}{|S|} \log \frac{|S_V|}{|S|}, \quad (7)$$

where A = candidate attribute, V = possible values of A , S = set of examples X , S_V = subset where $X_A = V$. In this method, the attribute with the highest gain ratio is selected as the splitting attribute.

2.2.2. Feature Subset Selection. CfsSubsetEval evaluator gives a subset of features which show high correlation with ground truth labels but are very less correlated to each other. In this research, we have performed both conventional search techniques and the newly introduced swarm search techniques [33]. Best first search is a conventional greedy search technique that finds subsets by backtracking and provides the flexibility of searching in forward, backward, and bidirections. Greedy stepwise search is similar to best first search as it also searches for subsets in a greedy manner, but instead of backtracking, it searches in a stepwise fashion and stops when a decrease of evaluation occurs after an addition or deletion of an attribute.

Swarm search methods are metaheuristics-based search methods that are nature-inspired optimization algorithms. Ant search [34] is a stochastic combinatorial optimization technique that is modelled on how the ant

colonies work. It finds solutions in the early stages of the search process. Bat search [35] is based on how bats determine object location by using sound waves. Bee search [36] is inspired by the foraging behaviour of honeybees. It solves optimization problems by doing exploitative neighbourhood search and random explorative search together. Cuckoo search [37] algorithm models the parasitic behaviour of cuckoos and the Levy flight behaviour of fruit flies while elephant search [38] models the behavioural characteristics of elephant herds. Firefly search [39] is inspired by the flashing patterns of fireflies and offers the advantage of automatic subdivision and the ability to deal with multimodality. Flower search [40] follows the survival of the fittest rule based on the process of pollination of flowers. Wolf search [41] models how wolves hunt for prey and survive by avoiding their enemies. It is known to give good results when used in large search spaces. Rhinoceros search [42] models the herding behaviour of rhinos and can be used for solving global continuous optimization problems.

2.3. Rule-Based Classifiers. Rule-based classifiers classify using “if... then... else” rules. If the set of feature values in a test sample conforms to a set of rules, then the sample is classified as belonging to the class associated with that set of rules. The classification rules can be built by directly extracting them either from the data (direct method) or from other classification models (indirect method). These classifiers are very expressive in their results as the rules learned by them can help understand the underlying mechanics of the classification. Some of the rules generated by one of our classifiers are discussed in Section 4.3.

We have applied the rule-based classifiers: PART and JRIP for bearing fault diagnosis. PART [43] is an efficient and accurate rule-based algorithm based on partial decision trees. It builds rules in a recursive fashion by using the “separate-and-conquer” strategy. It forms a rule and then removes the instances that rule covers and again creates a rule for remaining instances, and these steps repeat until the number of remaining samples becomes zero. JRIP [44], an implementation of the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm, builds association rules using reduced error pruning (REP). First, a rule is added with conjuncts to improve the information gain. Then, the rules in the set are pruned using incremental REP. This is repeated until the criteria for discretion length are achieved. The rule set is then optimized.

3. Methodology

The flowchart illustrating the steps used in our proposed methodology is shown in Figure 1. First, bearing vibration signals are bifurcated into nonoverlapping signals of length 1024 each. Empirical mode decomposition is applied on these signals to obtain their intrinsic mode functions. Second, each of the parent signals is then replaced by its IMF having the highest Pearson correlation with the parent signal, for further analysis

FIGURE 1: Flowchart of the proposed methodology.

TABLE 2: Pearson's correlation coefficient between a raw vibration signal and each of its IMFs

	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8	IMF9	IMF10	IMF11	IMF12
$x_1(n)$	0.7732	0.6790	0.5711	0.6099	0.6174	0.6927	0.7220	0.6656	0.6563	0.6770	0.5092	0.2464
$x_2(n)$	0.8925	0.6375	0.5286	0.5046	0.6102	0.6432	0.5587	0.5329	0.5329	0.3920	0.2470	0.1647

Note: $x_1(n)$ and $x_2(n)$ belong to IR 007 and BB 021, respectively.

TABLE 3: Dataset A for fault type detection.

Class	No. of train samples	No. of test samples
Normal	2649	665
Inner race	3774	961
Outer race	3473	799
Ball bearing	3752	987
Total	13648	3412

(Table 2). In this paper, we will refer to them as representative IMFs. Third, 30 statistical time-domain features (Table 1) are extracted from these IMFs. Fourth, an average ranking of all the extracted features is obtained using filter-based ranking methods. Fifth, filter-based feature subset selection methods are used to find subsets of features having low intracorrelation and high correlation with the truth labels. The results obtained from the fourth and fifth steps are analysed together to obtain a set of significant features. Finally, the selected feature subset is used to train rule-based classifiers and their ensembles to perform classification for both fault type detection (Table 3) and fault severity detection (Table 4). The two classification models are also trained on a modified feature subset, formed by replacing one of the features, namely, normal negative log likelihood of single Gaussian (T30), by its Gaussian mixture model counterpart (T30*). The T30* feature values have been computed for each vibration signal. The results obtained show the performance of classifiers to be slightly improved (Section 4.7, Exp. 6). The trained models from our pipeline are also tested on a bearing dataset obtained from the water pump at an operational power plant (Section 4.8, Exp. 7).

In this paper, EMD, representative IMF selection, and feature extraction have been performed using scripts written in MATLAB R2017a. The ranking and selecting of subset of features have been conducted through the open-source tool, Waikato Environment for Knowledge Analysis (Weka) [45]. Moreover, all the rule-based classifiers and their ensembles have been trained and tested in Weka. The Matlab scripts and models have been executed on an 8 GB RAM machine with Intel Core i5-7200U CPU having 2.50 GHz speed.

4. Experimental Results and Discussion

4.1. Dataset Preparation. In this paper, we have used the Case Western Reserve University Bearing Dataset. The setup (Figure 2) consists of a 2 hp electric motor, transducer/encoder, dynamometer, and accelerometers. Data for normal bearings were recorded using the deep groove ball bearing SKF6205-2RS JEM with dimensions 25 mm × 52 mm × 15 mm at drive end for motor loads of 0, 1, 2, and 3 horsepower (motor speeds of 1797, 1772, 1750, and

1730 rpm, respectively). Electrodischarge machining was used to induce faults of diameters: 0.007 inches, 0.014 inches, and 0.021 inches in the SKF6205 bearing and 0.028 inches in its NTN equivalent. After reinstalling these faulted bearings into the test motor, vibration data were recorded for the same motor loads. Vibration data were collected using a 16-channel DAT recorder for single-point drive-end defects at 12,000 samples/second.

The data consist of signals corresponding to four cases: outer raceway defect, inner raceway defect, rolling element, i.e., ball defect, and the normal. The defect of the outer raceway fault was located at 6 o'clock (orthogonal to the load zone). Data samples of size 1024 points were extracted from these vibration signals in a linear nonoverlapping fashion, i.e., for each signal, the first 1024 points were extracted as its first data sample, the next 1024 points as its second data sample, and so on. A total of 17060 samples consisting of 3314 normal, 4735 inner race, 4272 outer race, and 4739 ball data samples were extracted. Each of the resulting 17060 signals, of 1024 data points, was decomposed by EMD resulting in 12–17 IMFs along with residue. The IMF having the highest Pearson correlation with the parent signal was taken as its representative signal since it retains most of the information of the original signal (Figure 3). The correlation coefficient $\rho_{X,Y}$ was computed as follows:

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \quad (8)$$

where E is the expectation, σ_X and σ_Y are standard deviations of X and Y , and μ_X and μ_Y are means of X and Y , respectively. The correlation coefficient values obtained for two sample vibration signals and their IMFs are shown in Table 2. The representative signals of all the parent signals were then used to extract 30 statistical time-domain features listed in Table 1. Min-max normalization was applied to scale the features. The values of every feature were transformed into a decimal between 0 and 1 (0,1 both inclusive) by applying the following equation:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \quad (9)$$

Two kinds of datasets were prepared: dataset A for fault type detection and dataset B for fault severity detection. Dataset A has the data labelled into 4 classes—outer raceway defect, inner raceway defect, ball bearing defect, and normal whereas dataset B has the data labelled into 12 classes defined by fault types and fault sizes (in inches). Both the datasets were divided into train and test sets using 80:20 split. The number of samples in train and test sets for dataset A and dataset B is shown in Tables 3 and 4 respectively.

TABLE 4: Dataset B for fault severity detection.

Class	Fault type	Fault size	No. of train samples	No. of test samples
Normal	—	—	2663	651
IR 007	Inner race	0.007"	1135	292
IR 014	Inner race	0.014"	1139	277
IR 021	Inner race	0.021"	1152	269
IR 028	Inner race	0.028"	373	98
OR 007	Outer race	0.007"	1158	267
OR 014	Outer race	0.014"	1126	296
OR 021	Outer race	0.021"	1134	291
BB 007	Ball bearing	0.007"	1131	288
BB 014	Ball bearing	0.014"	1136	289
BB 021	Ball bearing	0.021"	1125	299
BB 028	Ball bearing	0.028"	376	95
Total			13648	3412

FIGURE 2: CWRU test rig schematic diagram.

FIGURE 3: Plots of the best IMFs selected on the basis of Pearson correlation coefficient. A drive-end sample of size 1024 points from each of the 12 classes has been depicted.

TABLE 5: Feature ranking of statistical features for dataset A.

Feature	Feature description	Chi-squared statistics	Info gain	Gain ratio	Average rank score
T22	Zero crossing rate	1	1	1	1
T27	Hjorth parameter—mobility	2	2	3	2.33
T30	Normal negative log likelihood	3	3	4	3.33
T26	Hjorth parameter—activity	4	4	7	5
T2	Root mean square (RMS)	6	6	5	5.67
T6	Standard deviation	5	5	8	6
T19	Root-sum-of-squares (RSSQ)	7	7	6	6.67
T18	Geometric mean	12	12	2	8.67
T20	Mean absolute deviation	9	9	9	9
T23	Entropy	8	8	14	10
T29	Weibull negative log likelihood	10	11	12	11
T3	Root	11	10	13	11.33
T21	Median absolute deviation	13	13	11	12.33
T9	75th percentile	16	15	15	15.33
T17	Kurtosis factor	15	14	17	15.33
T24	Histogram upper bound	14	16	16	15.33
T8	25th percentile	20	19	10	16.33
T5	Peak-to-peak	17	18	19	18
T4	Max	18	17	20	18.33
T25	Histogram lower bound	19	20	18	19
T28	Hjorth parameter—complexity	21	21	21	21
T16	Skewness factor	22	22	22	22
T15	Clearance factor	23	23	23	23
T11	Kurtosis	24	24	24	24
T10	Skewness	25	26	25	25.33
T12	Crest factor	27	25	26	26
T13	Shape factor	26	27	27	26.67
T14	Impulse factor	28	28	28	28
T7	Median	29	29	29	29
T1	Mean	30	30	30	30

TABLE 6: Feature ranking of statistical features for dataset B.

Feature	Feature description	Chi-squared statistics	Info gain	Gain ratio	Average rank score
T2	Root mean square (RMS)	3	4	4	3.67
T18	Geometric mean	4	6	2	4.00
T20	Mean absolute deviation	1	3	8	4.00
T30	Normal negative log likelihood	2	5	9	5.33
T22	Zero crossing rate	15	2	1	6.00
T27	Hjorth parameter—mobility	16	1	3	6.67
T17	Kurtosis factor	5	13	5	7.67
T19	Root-sum-of-squares (RSSQ)	6	14	6	8.67
T9	75th percentile	7	15	7	9.67
T23	Entropy	8	12	11	10.33
T26	Hjorth parameter—activity	10	7	16	11.00
T29	Weibull negative log likelihood	11	8	15	11.33
T6	Standard deviation	12	9	14	11.67
T21	Median absolute deviation	13	10	12	11.67
T8	25th percentile	9	16	10	11.67
T3	Root	14	11	13	12.67
T24	Histogram upper bound	17	17	19	17.67
T4	Max	18	18	18	18.00
T5	Peak-to-peak	19	19	17	18.33
T25	Histogram lower bound	20	20	20	20.00
T28	Hjorth parameter—complexity	22	21	22	21.67
T11	Kurtosis	23	22	21	22.00
T16	Skewness factor	21	23	24	22.67
T15	Clearance factor	24	24	23	23.67
T12	Crest factor	25	25	25	25.00
T7	Median	26	26	27	26.33
T1	Mean	27	27	26	26.67
T10	Skewness	28	28	28	28.00
T14	Impulse factor	29	29	30	29.33
T13	Shape factor	30	30	29	29.67

4.2. Exp. 1: Feature Ranking and Feature Subset Selection.

Three feature ranking techniques, based on the statistical metrics of chi-squared, info gain, and gain ratio, were applied to the data. These techniques ranked the features according to their importance in classifying the samples into 4 classes for dataset A and likewise into 12 classes for dataset B. The threshold for the search method used by the feature ranking techniques was set in a way to ensure none of the features gets discarded. The ranking of the features is shown in Tables 5 and 6. The 3rd, 4th, and 5th columns list the rank obtained by ranker methods chi-squared, info gain, and gain ratio, respectively, for the feature in the corresponding row. The last column shows the average rank score calculated by averaging out the ranks of features over all the three ranking metrics.

The evaluator named CfsSubsetEval was used for further validation. This evaluator gives a subset of features which show high correlation with ground truth labels but low correlation to each other. 11 search techniques, including both conventional search techniques (best first search and greedy stepwise search) and the newly introduced swarm search techniques, were performed. Tables 7 and 8 show the results obtained with these techniques for the two datasets.

For implementing best first search, bidirectional search was used, while in greedy stepwise, forward search was used. The number of particles in the swarm, chaotic coefficient, and mutation probability were set as 20, 4.0, and 0.01, respectively, for all searches. In bee search, the discount factor for crossover was set as 0.8, and in cuckoo search, the sigma rate was kept as 0.69. For firefly search, zero distance attractiveness for the firefly population was set as 0.33. The same value was taken for wolf search. In flower search, a pollination rate of 0.33 was used.

From Table 7, it is observed that the subset of root mean square (T2), geometric mean (T18), rate of zero crossing (T22), Hjorth parameter—mobility (T27), and normal negative log likelihood for single Gaussian (T30) is selected by majority of search techniques. It is also observed that these 5 features are among the top 10 average rank scorers as highlighted in Table 5, hence validating the importance of these 5 features for fault type diagnosis. It can be noted that the features standard deviation (T6), RSSQ (T19), and Hjorth parameter—activity (T26) are ranked well by the ranking metrics but get selected by only a few subset search techniques. This could be due to the way subset selection works. Since one of the criteria for a good subset is that the subset elements should have very less correlation among each other, features like T6, T19, and T26 got rejected by most of the search techniques as they may be showing some correlation with an already chosen member of the set. Since our aim is to select the best performing subset for fault diagnosis, we selected the set of only the 5 features stated above.

For fault severity detection, from Table 8, we observe that the subset of features—root mean square (T2), RSSQ (T19), kurtosis factor (T17), geometric mean (T18), rate of zero crossing (T22), Hjorth parameter—mobility (T27), normal negative log likelihood for single Gaussian (T30), and 75th percentile (T9)—is selected by majority of search techniques, where these 8 features are also among the top 10 average rank scorers in Table 6.

TABLE 7: Feature subset selection results for dataset A.

Technique	Set of features selected
Best first	T2, T18, T22, T27, T30
Greedy stepwise search	T2, T18, T22, T27, T30
Ant search	T2, T18, T22, T27, T30
Bat search	T18, T22, T26, T27, T30
Bee search	T18, T19, T22, T27, T30
Cuckoo search	T2, T18, T22, T27, T30
Elephant search	T18, T22, T26, T27, T30
Firefly search	T6, T22, T26, T27, T30
Flower search	T18, T19, T22, T27, T30
Rhinoceros search	T2, T18, T22, T27, T30
Wolf search	T2, T18, T22, T27, T30

TABLE 8: Feature subset selection results for dataset B.

Technique	Set of features selected
Best first	T2, T9, T17, T18, T19, T22, T27, T30
Greedy stepwise	T2, T9, T17, T18, T19, T22, T27, T30
Ant search	T2, T9, T17, T18, T19, T22, T27, T30
Bat search	T9, T17, T18, T22, T26, T27, T30
Bee search	T2, T17, T18, T21, T22, T27, T30
Cuckoo search	T9, T19, T18, T22, T26, T27, T29, T30
Elephant search	T2, T9, T17, T19, T22, T27, T30
Firefly search	T17, T18, T20, T21, T22, T26, T27
Flower search	T2, T15, T17, T18, T22, T27, T19
Rhinoceros search	T2, T3, T9, T19, T22, T27, T30
Wolf search	T9, T17, T18, T20, T22, T27, T30

4.3. Exp. 2: Fault Type Detection with Feature Selection.

Rule-based classifiers, PART and JRIP, were used with ensemble techniques boosting (multiboost [46] in our case) and bagging for performing fault type detection of the 4 classes—normal, inner race, outer race, and ball bearing. The ensemble techniques make use of a single learning algorithm to train multiple models on the dataset and then average out these models to get a final prediction. While bagging helps in getting an ensemble model with less variance by assigning equal weights to individual models for averaging, boosting works by reducing both variance and bias. Both of them help in increasing the stability of the models.

For training PART, the value of pruning confidence factor was taken as 0.25, while in JRIP, the minimum of the instances in a rule was taken as 2 and two optimization runs were performed. These classifiers were trained on the reduced dataset A obtained by retaining only the 5 features that were found to be significant, as discussed in the previous section. The data augmentation achieved by the initial splitting of vibration signals into signals of 1024 length resulted in a large enough dataset, removing the need to perform cross-validation while training the classifiers. For implementing bagging and multiboost, 10 iterations were performed with a batch size of 100.

Table 9 reports the performance of the classifiers in terms of accuracy obtained on the train and test datasets. Two cases were considered as follows:

TABLE 9: Classifier performance in terms of train accuracy and test accuracy values for Exp. 2.

Classifier	Train accuracy		Test accuracy	
	T30	T30*	T30	T30*
PART	98.11	98.36	95.43	95.43
PART + multiboost	100	100	96.51	96.75
PART + bagging	98.72	98.81	96.49	96.72
JRIP	96.15	96.41	94.28	94.37
JRIP + multiboost	99.37	99.99	96.22	96.37
JRIP + bagging	98.10	98.21	96.04	96.25

- (a) When the selected feature set contains T2, T18, T22, T27, and T30 corresponding to column 1 of Table 9
- (b) When T30 is replaced with its Gaussian mixture model equivalent T30*, the feature set in this case being T2, T18, T22, T27, and T30* corresponding to column 2 of Table 9

This section focuses on discussing the results for the case (a). The results for case (b) are discussed later in Section 4.7. It is observed that PART when used with multiboost outperforms all the 5 classifiers and gives the best performance with an accuracy of 96.51% on the test dataset and 100% on the train dataset. The performance of the features is further studied by inspecting the rules generated by them. Rules are propositional formulae consisting of one or more atomic formulae (enclosed in parentheses as shown below) that are joined together using logical connectives AND $\wedge$ and IMPLIES $\rightarrow$. The number at the end of a rule indicates the number of instances correctly classified by that rule. PART with multiboost generates a total of 87 rules from the 5 important features. With just a small number of rules, the entire training data of 13648 samples are fully and accurately classified. Some of the rules generated are listed as follows:

$$\begin{aligned}
 R1: & (T22 > 0.358398) \wedge (T27 \leq 0.318614) \wedge (T18 > 0.025684) \wedge (T2 \leq 0.477539) \rightarrow \text{inner race (660)}, \\
 R2: & (T22 > 0.416992) \wedge (T18 > 0.439639) \rightarrow \text{ball bearing (375)}, \\
 R3: & (T22 > 0.416992) \wedge (T18 > 0.086578) \wedge (T2 > 0.333767) \wedge (T18 \leq 0.177064) \rightarrow \text{outer race (350)}, \\
 R4: & (T2 \leq 0.050631) \wedge (T22 > 0.291016) \wedge (T22 \leq 0.357422) \rightarrow \text{normal (1635)}.
 \end{aligned} \tag{10}$$

From the above rules, it can be observed that a total of 1635 out of 2649 dataset instances were correctly classified as normal using just one rule R4 generated by the features T2 and T22. Each of the three remaining rules also classifies significant portions of samples in the dataset, thus showing the good performance of the features forming these rules. The rules are also very simple in structure, for instance, R2 constructed with the help of just 2 atomic formulae correctly classifies 375 ball bearing instances. The ability to easily distribute the samples into the 4 classes using simple rules shows the efficiency of the selected time-domain features.

Since the number of samples belonging to normal class is 2649, which is less than the number of samples in each of the faulty classes (around 3500 for each), our dataset is slightly class unbalanced. Hence, evaluating the performance of the rule-based classifiers on accuracy alone may not be comprehensive, and therefore, the performance of the classifiers on the test dataset has been further studied by calculating 2 other metrics: *F*-measure and geometric mean (*G*-mean). *F*-measure per class is calculated as the harmonic mean of precision and sensitivity. *G*-mean is the geometric mean of specificity and sensitivity. Table 10 reports the performance of the classifiers on dataset A using *F*-measure and *G*-mean scores. The *F*-measure and *G*-mean scores for all the classifiers are significantly high; in particular, PART + multiboost classifier outperforms the other classifiers. Hence, our rule-based classifiers are able to classify our slightly unbalanced dataset very well, in turn showing the good performance of the selected features.

To further validate the good performance of the selected features in fault type detection, confusion matrices were

built. Figure 4(a) shows the confusion matrix obtained on the test dataset with T30 parameter for PART with multiboost. The normal instances were accurately classified (with only 2 misclassifications), further confirming that they are well classified by our 5 important features even though normal class is slightly underrepresented. Another important observation is that most of the misclassifications involve ball bearing class. 22 instances each of inner race and outer race got misclassified as ball bearing. Also, a total of 56 ball bearing instances were incorrectly classified. To validate these observations, 3 features—T22, T27, and T30—were plotted pairwise using scatter plots as shown in Figure 5. It can be seen that for normal class, the plots show mostly nonoverlapping clusters, hereby confirming that the selected features are playing a very good role in classifying fault vs. no fault. The plots also show a lot of overlapping between ball bearing, inner race, and outer race clusters, confirming our observation about ball bearing class. This indicates a lot of diversity in the ball bearing instances that are unhandled by our features. Hence, features which can accurately distinguish this class from other classes are needed. A modification in one of the features resulting in a better classification for ball bearing instances has been discussed in Section 4.7.

4.4. Exp. 3: Fault Type Detection without Feature Selection. Table 11 reports the performance of the classifiers in fault type detection in terms of accuracy obtained on dataset A without feature selection. These classifiers perform better

TABLE 10: Classifier performance in terms of F -measure and G -mean values for Exp. 2.

Classifier	F -measure				Geometric mean (G -mean)			
	Normal	Inner race	Outer race	Ball bearing	Normal	Inner race	Outer race	Ball bearing
PART	0.990	0.955	0.949	0.932	0.994	0.970	0.969	0.949
PART + multiboost	0.995	0.967	0.958	0.948	0.997	0.977	0.974	0.962
PART + bagging	0.994	0.966	0.957	0.946	0.995	0.976	0.973	0.961
JRIP	0.987	0.931	0.929	0.934	0.992	0.958	0.947	0.951
JRIP + multiboost	0.993	0.965	0.955	0.944	0.995	0.972	0.973	0.961
JRIP + bagging	0.991	0.958	0.957	0.944	0.995	0.971	0.973	0.960

FIGURE 4: Confusion matrices for Exp. 2 using multiboost + PART on the test dataset (a) with T30 and (b) with T30*.

than their feature selected counterparts, in terms of accuracies obtained. Again, in terms of overall performance, PART + multiboost outperforms all the 5 classifiers, giving the highest train and test accuracies of 100% and 99.13%, respectively. On comparing these with the train and test accuracies of 100% and 96.51% obtained with feature selection, we observe that, on the train data, the selected features were alone sufficient for attaining a perfect accuracy, while in the case of test data too, a major contribution in achieving the high accuracy can be credited to the selected features.

4.5. Exp. 4: Fault Severity Detection with Feature Selection.

This section detects the severity of fault types by performing a 12-class classification using rule-based classifiers on the selected eight features obtained for dataset B. The 12 labels used for severity detection are shown in Table 4 where the severity of a fault has been characterised by its size in inches. Table 12 shows the results obtained wherein, once again PART, when used with multiboost, outperforms all the 5 classifiers and gives the best performance accuracy of 100% on the train dataset and 97.92% on the test dataset. The

results obtained with the parameter T30* are discussed later in Section 4.7.

Having few IR 028 and BB 028 samples and almost thrice the count for other severity labels, the dataset is unbalanced classwise, and hence, we obtained the F -measure and G -mean values to further study the performance of the selected features. Table 13 reports the F -measure and G -mean values obtained with PART + multiboost for fault severity detection. We observe that both the underrepresented classes and the remaining classes are accurately classified. Hence, the selected features for fault severity detection are performing well despite the data being class unbalanced.

4.6. Exp. 5: Fault Severity Detection without Feature Selection.

In this experiment, we fed dataset B with all the 30 statistical features to the rule-based classifiers to again study fault severity detection, results for which are shown in Table 14. In comparison with Table 12 data, the highest train accuracy of 100% remains unaltered while the test accuracy sees a gain of 0.76%. The perfect train accuracy and the small gain in test accuracy show the superior performance of our selected eight features over the other features, confirming their large

FIGURE 5: Pairwise scatter plots for the top 3 ranked features. (a) Between T27 and T30. (b) Between T22 and T27. (c) Between T22 and T30. Here T22, T27, and T30 are rate of zero crossing, Hjorth parameter—mobility, and normal negative log likelihood for single Gaussian, respectively.

contribution towards the accurate classification results. This is in line with our observation for the selected features for fault type detection in Experiment 3.

4.7. Exp. 6: Replacing the Normal Negative Log Likelihood for Single Gaussian Parameter with Its Gaussian Mixture Model Variant for Exp. 2–5. In Experiment 2, the feature T30,

normal negative log likelihood for single Gaussian, comes in the set of important features for both fault type and fault severity detection. For instance, in the case of fault type detection, it occupies rank 3 in the list of average ranks given by feature ranking methods and is selected by all the 5 feature subset methods. This feature gives the negative of the likelihood that a normal distribution of given mean and standard deviation fits the sample data, in our case, the

TABLE 11: Classifier performance in terms of train accuracy and test accuracy values for dataset A for Exp. 3.

Classifier	Train accuracy		Test accuracy	
	T30	T30*	T30	T30*
PART	98.63	98.82	96.04	96.25
PART + multiboost	100	100	99.13	99.63
PART + bagging	99.47	99.8	98.6	98.81
JRIP	97.71	97.82	96.6	95.96
JRIP + multiboost	99.8	99.99	97.04	97.27
JRIP + bagging	98.82	98.9	97.04	97.25

TABLE 12: Classifier performance in terms of train accuracy and test accuracy values for Exp. 4.

Classifier	Train accuracy		Test accuracy	
	T30	T30*	T30	T30*
PART	96.1	96.14	94.42	95.18
PART + multiboost	100	100	97.92	98.56
PART + bagging	97.9	97.96	97.41	97.53
JRIP	91.22	91.83	90.99	91.9
JRIP + multiboost	96.1	96.29	96.03	96.15
JRIP + bagging	95.63	95.94	95.35	95.82

sample data being the representative IMF of bearing vibration signal. However, on selecting some random samples from the set of representative IMF signals and plotting their frequency distribution histograms with fitting curves, shown in Figure 6, we found that the idea of fitting a single Gaussian curve will not work well with some of the signals. Some signals, like samples 1 and 2 in Figures 6(a) and 6(b), respectively, have more than one distinct maxima, and a single Gaussian curve will not be able to accurately model these signals. This trend was mostly seen for samples belonging to faulty classes. Hence, we replaced the T30 feature with its Gaussian mixture model variant, calling it T30*. The normal negative log likelihood for Gaussian mixture model (T30*) gives the negative of the likelihood that a GMM of k means and k standard deviations fits the sample data. Mathematical formula for normal negative log likelihood of GMM can be defined as follows:

$$-\log \Pr(\pi, \mu, \Sigma) = -\sum_{i=1}^N \log \sum_{k=1}^K \pi_k N\left(x_i | \mu_k, \Sigma_k\right), \quad (11)$$

where π_k is the mixing coefficient of the k th Gaussian. Here, we instantiated k by the number of local maxima in the probability distribution of the sample data. Table 15 reports the normal negative log likelihood for GMM values obtained for the 3 samples shown in Figure 6. The normal negative log likelihood for GMM for Sample 1 is -466 , which is far better than its normal negative log likelihood for single Gaussian equal to -143.347 , in terms of fitting of the model. The same is the case with Sample 2. Sample 3 shows a very slight change in value because it can be fitted well with a single Gaussian.

Tables 9 and 11 show the results obtained for fault type detection when T30 was taken as one of the 5 selected features and when it was replaced by T30*. According to

Table 9, PART with multiboost gives an accuracy of 100% on the train dataset for both T30 and T30* and shows a slight jump in test accuracy from 96.51% with T30 to 96.75% with T30*. For the rest of the classifiers, we see a slight improvement in train and test accuracy with T30* over T30. From the two confusion matrices in Figure 4, the number of true positives for the ball bearing class has increased from 931 with T30 to 939 with T30*, accompanied by a decrease in the misclassified instances of ball bearing into inner race and outer race classes from 24 to 22 and 28 to 22, respectively. This further confirms the better performance of T30* as compared to T30. Similarly, the accuracy values obtained in Table 11 show better accuracy values with parameter T30* than the values obtained with T30 for all the classifiers. This again shows the effectiveness of T30* over T30.

A similar trend is observed for fault severity detection. From Table 12, it is seen that on comparing T30 with T30*, the train accuracy remains at its peak of 100% while test accuracy increases by 0.64% to become 98.56%. T30* performs better than T30 in the case of without feature selection as well, as can be seen from Table 14. A more clear picture is provided by the confusion matrices in Figure 7 where we see that for T30, the misclassifications are more than that for T30*. We believe that for bearing datasets having a majority of vibration signals with multiple peaks, this feature can prove to be a key performer because of its ability to represent such signals well. Hence, normal negative log likelihood of GMMs has the potential of being a more accurate and reasonable parameter for feature selection and bearing fault diagnosis than the normal negative log likelihood of single Gaussian.

4.8. Exp. 7: Testing Our Methodology on the Bearing Data of an Operating Power Plant. In this experiment, we have used bearing data from 500 MW Kosti Thermal Power Plant commissioned by Bharat Heavy Electricals Limited, India. The setup, shown in Figure 8, consists of a 7 hp vertical electric motor of service water pump. Data for normal baseline bearings were recorded using the deep groove ball bearing SKF6303-2Z/C3 with dimensions 17 mm $\times$ 47 mm $\times$ 14 mm at drive end with a motor speed of 3000 rpm. Vibration data were collected at 12,000 samples/second. Vibration data are taken for 2 different timestamps, and for every timestamp, data have been taken for both vertical and horizontal positions, leading to a total of 4 vibration signals. We performed the same data preprocessing, including bifurcating the signals to subsignals of length 1024, EMD, and feature extraction, on these samples as done for the CWRU dataset.

Since the bearing configuration matches with that used in CWRU dataset, we mixed these test samples with the 20% test set separated from the CWRU dataset A. The trained model of PART + multiboost for fault type detection with feature selection, obtained in Exp. 2, was chosen for labelling instances as outer raceway defect, inner raceway defect, ball defect, and the normal. The overall accuracy and average F -measure achieved were 96.75% and 0.972, respectively. Furthermore, per instance classification results were obtained and it was observed that all the normal samples taken

TABLE 13: PART + multiboost performance in terms of F -measure and G -mean values for Exp. 4.

	Normal	IR 007	IR 014	IR 021	IR 028	OR 007	OR 014	OR 021	BB 007	BB 014	BB 021	BB 028
F -measure	0.998	0.988	0.982	0.980	1.000	0.998	0.965	0.981	0.962	0.951	0.953	1.000
G -mean	0.999	0.994	0.987	0.993	1.000	1.000	0.984	0.992	0.976	0.972	0.968	1.000

TABLE 14: Classifier performance in terms of train accuracy and test accuracy values for Exp. 5.

Classifier	Train accuracy		Test accuracy	
	T30	T30*	T30	T30*
PART	98.01	98.15	95.68	95.71
PART + multiboost	100	100	98.68	99.12
PART + bagging	99.36	99.39	97.5	97.56
JRIP	93.92	94.61	93.77	94.21
JRIP + multiboost	98.78	98.98	96.32	96.59
JRIP + bagging	97.73	97.79	96.08	96.35

FIGURE 6: Frequency distribution histograms with fitting curves for selected samples. (a) Sample 1. (b) Sample 2. (c) Sample 3.

TABLE 15: Comparison of normal negative log likelihood for Gaussian mixture models with normal negative log likelihood for single Gaussian for selected samples.

Sample no.	Normal negative log likelihood for single Gaussian	Normal negative log likelihood for GMM
1	-143.347	-466.000
2	-567.550	-987.950
3	901.479	901.479

The values shown are taken before min-max scaling was applied.

FIGURE 7: Confusion matrices for Exp. 4 using PART + multiboost on the test dataset (a) with T30 and (b) with T30*.

FIGURE 8: Experimental setup consisting of vertical electric motor of service water pump at Kosti Thermal Power Plant.

from the power plant were correctly classified. Hence, the results were reasonably good. The mixing of power plant samples with CWRU test samples ensured that the good accuracy was not due to any bias shown by the classifier towards the normal class, thereby also leading to a good F -measure value. Furthermore, when T30 was replaced by T30* for this dataset, we did not see any improvements in the accuracy and F -measure values. This might be due to the mostly single Gaussian nature of the newly added normal vibration signals taken from the power plant.

The following points discuss and summarize the obtained results:

- (1) For the 4-class fault type detection problem, a subset of the features—root mean square (T2), geometric mean (T18), rate of zero crossing (T22), Hjorth

parameter mobility (T27), and normal negative log likelihood for Gaussian mixture model (T30*)—performs well with rule-based classifiers, giving the maximum test accuracy of 96.75% obtained with PART + multiboost.

- (2) For the 12-class fault severity estimation problem, a subset of the features—root mean square (T2), 75th percentile (T9), kurtosis factor (T17), geometric mean (T18), RSSQ (T19), rate of zero crossing (T22), Hjorth parameter mobility (T27), and normal negative log likelihood for single Gaussian (T30*)—performs very well, producing the maximum test accuracy of 98.56% with PART + multiboost.
- (3) An improvement in classification accuracy is observed when normal negative log likelihood for

single Gaussian (T30) is replaced by its GMM variant (T30*). This is because some signals have more than one distinct maxima and a single Gaussian curve cannot model these signals accurately. Thus, instead of a single Gaussian curve, it is better to fit these signals with a Gaussian mixture model.

- (4) The *F*-measure and *G*-Mean scores for all the rule-based classifiers are significantly high, confirming the potential of the chosen features to accurately classify the unbalanced dataset.
- (5) On comparing the results with feature selection vs. without feature selection, the latter performs better for both fault diagnosis and fault severity estimation giving 99.63% and 99.12% test accuracies, respectively. Despite this observation, the good accuracy values obtained with the selected features show that a major contribution in achieving the high accuracy in the case of without feature selection can be credited to the selected features. Hence, these selected features play a major role in the diagnostic ability of the classifiers.
- (6) The efficacy of the selected time-domain features is further confirmed by their good performance in classifying samples taken from an operating power plant.

5. Conclusions

So far, research in bearing fault diagnosis and severity estimation has majorly focussed on improving the performance of classification algorithms. However, the applicability of these approaches mainly depends on the quality of features extracted from the bearing signals. In light of this, this paper focussed instead on optimising the input features fed to these classification systems. We believe that if new input features are proposed and/or the input features are selected comprehensively using an in-depth analysis, they can contribute more in bearing fault diagnosis, covering all types of fault samples in an exhaustive manner. This can in turn lead to the existing classification methods to perform much better. In an attempt to achieve this, this paper proposed the usefulness of Hjorth parameters and normal negative log likelihood for Gaussian mixture model (GMM) for rolling element ball bearing fault detection and severity estimation. An in-depth two-step approach, based on filter-based ranking with 3 metrics and feature subset selection with 11 search techniques, was developed to select an optimal feature subset for bearing fault type detection and severity estimation. Filter-based selection techniques being model-agnostic make sure that the selected features work efficiently with any machine learning classifier. In this paper, we used rule-based classifiers and their ensembles for validation of our results in terms of diagnostic accuracy.

The obtained results show that features root mean square, geometric mean, rate of zero crossing, Hjorth parameter—mobility, and normal negative log likelihood are

important statistical features for bearing fault diagnosis. For fault severity detection, these five parameters and three additional parameters—75th percentile, kurtosis factor, and RSSQ, collectively form an important and useful feature subset. We also proposed a novel improvement in the performance of one of the selected high ranked features, normal negative log likelihood for single Gaussian, by suggesting its replacement with its Gaussian mixture model (GMM) variant. This is because some vibration signals have more than one distinct maxima and a single Gaussian curve cannot model these signals accurately. This also leads us to conclude that existing features could be improvised to achieve maximum possible coverage of the underlying data.

The performance of the feature set was validated by running experiments for different conditions—with and without feature selection—and comparing their results. The dataset being class unbalanced proved that the good performance of the feature set is independent of the class distributions of the dataset. Since in real-world scenarios one can never exactly replicate the experimental conditions of a class-balanced dataset, our results guarantee that the selected features are independent of class-balanced scenarios. In addition to CWRU test rig dataset, the performance was further validated on a bearing dataset obtained from an operational power plant. The results show that the methodology is robust, aligns well with the real-world situation of class unbalanced datasets, and achieves good performance for both fault diagnosis and fault severity estimation. We also believe that the results obtained from this research can pave the way for researchers to change their current method of developing diagnostic systems by first analysing the feature set comprehensively and then shifting their focus to the next step of classification. For future studies, the selected high-performing statistical features obtained in this paper can be combined with features extracted from different domains to build a more robust multidomain feature set. This work can also be extended by considering different types of bearing specifications. This will prevent the diagnostic system from becoming dependent on the vibration dataset of a specific bearing.

Data Availability

The Case Western Reserve University (CWRU) Bearing Dataset is available in [30]. The bearing data taken from the operational power plant were taken for academic research and are not publicly available.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors are thankful to Case Western Reserve University and Bharat Heavy Electricals Limited for providing them access to their ball bearing datasets.

References

- [1] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—reviews, methodology and applications," *Mechanical Systems and Signal Processing*, vol. 42, no. 1-2, pp. 314–334, 2014.
- [2] A. Rai and S. H. Upadhyay, "A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings," *Tribology International*, vol. 96, pp. 289–306, 2016.
- [3] A. Kumar and R. Kumar, "Role of signal processing, modeling and decision making in the diagnosis of rolling element bearing defect: a review," *Journal of Nondestructive Evaluation*, vol. 38, no. 1, 2019.
- [4] M. Zeng, Y. Yang, J. Zheng, and J. Cheng, "Maximum margin classification based on flexible convex hulls for fault diagnosis of roller bearings," *Mechanical Systems and Signal Processing*, vol. 66-67, pp. 533–545, 2016.
- [5] L. Lu, J. Yan, and C. W. de Silva, "Dominant feature selection for the fault diagnosis of rotary machines using modified genetic algorithm and empirical mode decomposition," *Journal of Sound and Vibration*, vol. 344, pp. 464–483, 2015.
- [6] R. K. Patel and V. K. Giri, "Feature selection and classification of mechanical fault of an induction motor using random forest classifier," *Perspectives in Science*, vol. 8, pp. 334–337, 2016.
- [7] O. Janssens, V. Slavkovikj, B. Vervisch et al., "Convolutional neural network based fault detection for rotating machinery," *Journal of Sound and Vibration*, vol. 377, pp. 331–345, 2016.
- [8] P. K. Kankar, S. C. Sharma, and S. P. Harsha, "Fault diagnosis of ball bearings using machine learning methods," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1876–1886, 2011.
- [9] X. Wang, Y. Zheng, Z. Zhao, and J. Wang, "Bearing fault diagnosis based on statistical locally linear embedding," *Sensors*, vol. 15, no. 7, pp. 16225–16247, 2015.
- [10] M. D. Prieto, G. Cirrincione, A. G. Espinosa, J. A. Ortega, and H. Henao, "Bearing fault detection by a novel condition-monitoring scheme based on statistical-time features and neural networks," *IEEE Transactions on Industrial Electronics*, vol. 60, p. 3398, 2012.
- [11] D. Dou and S. Zhou, "Comparison of four direct classification methods for intelligent fault diagnosis of rotating machinery," *Applied Soft Computing*, vol. 46, pp. 459–468, 2016.
- [12] X. Yan and M. Jia, "A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing," *Neurocomputing*, vol. 313, pp. 47–64, 2018.
- [13] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 108–126, 2013.
- [14] Q.-q. Chen, S.-w. Dai, and H.-d. Dai, "A rolling bearing fault diagnosis method based on EMD and quantile permutation entropy," *Mathematical Problems in Engineering*, vol. 2019, Article ID 3089417, 8 pages, 2019.
- [15] S. Cho, M. R. Shahriar, and U. Chong, "Identification of significant intrinsic mode functions for the diagnosis of induction motor fault," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. EL72–EL77, 2014.
- [16] H. Han, S. Cho, S. Kwon, and S.-B. Cho, "Fault diagnosis using improved complete ensemble empirical mode decomposition with adaptive noise and power-based intrinsic mode function selection algorithm," *Electronics*, vol. 7, no. 2, p. 16, 2018.
- [17] B. Sreejith, A. Verma, and A. Srividya, "Fault diagnosis of rolling element bearing using time-domain features and neural networks," in *Proceedings of the 2008 IEEE Region 10 and the 3rd International Conference on Industrial and Information Systems*, IEEE, Kharagpur, India, 2008.
- [18] S. Abbasion, A. Rafsanjani, A. Farshidianfar, and N. Irani, "Rolling element bearings multi-fault classification based on the wavelet denoising and support vector machine," *Mechanical Systems and Signal Processing*, vol. 21, no. 7, pp. 2933–2945, 2007.
- [19] P. E. William and M. W. Hoffman, "Identification of bearing faults using time domain zero-crossings," *Mechanical Systems and Signal Processing*, vol. 25, no. 8, pp. 3078–3088, 2011.
- [20] K. A. Korba and F. Arbaoui, "SVM multi-classification OD induction machine's bearings defect using vibratory analysis based on empirical mode decomposition," *International Journal of Applied Engineering Research*, vol. 13, p. 6579, 2018.
- [21] V. Sugumaran and K. I. Ramachandran, "Fault diagnosis of roller bearing using fuzzy classifier and histogram features with focus on automatic rule learning," *Expert Systems with Applications*, vol. 38, no. 5, pp. 4901–4907, 2011.
- [22] C. Grover and N. Turk, "Rolling element bearing fault diagnosis using empirical mode decomposition and Hjorth parameters," *Procedia Computer Science*, vol. 167, pp. 1484–1494, 2020.
- [23] J. B. Ali, N. Fnaiech, L. Saidi, B. Chebel-Morello, and F. Fnaiech, "Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals," *Applied Acoustics*, vol. 89, p. 16, 2015.
- [24] Y. Fan, C. Zhang, Y. Xue, J. Wang, and F. Gu, "A bearing fault diagnosis using a support vector machine optimised by the self-regulating particle swarm," *Shock and Vibration*, vol. 2020, Article ID 9096852, 11 pages, 2020.
- [25] Y. Chen, T. Zhang, Z. Luo, and K. Sun, "A novel rolling bearing fault diagnosis and severity analysis method," *Applied Sciences*, vol. 9, no. 11, p. 2356, 2019.
- [26] A. Kumar, C. P. Gandhi, Y. Zhou, H. Tang, and J. Xiang, "Fault diagnosis of rolling element bearing based on symmetric cross entropy of neutrosophic sets," *Measurement*, vol. 152, Article ID 107318, 2020.
- [27] A. Kumar, C. P. Gandhi, Y. Zhou, R. Kumar, and J. Xiang, "Variational mode decomposition based symmetric single valued neutrosophic cross entropy measure for the identification of bearing defects in a centrifugal pump," *Applied Acoustics*, vol. 165, Article ID 107294, 2020.
- [28] D. Dou, J. Jiang, Y. Wang, and Y. Zhang, "A rule-based classifier ensemble for fault diagnosis of rotating machinery," *Journal of Mechanical Science and Technology*, vol. 32, no. 6, pp. 2509–2515, 2018.
- [29] Y. Li, J. Wang, L. Duan et al., "Association rule-based feature mining for automated fault diagnosis of rolling bearing," *Shock and Vibration*, vol. 2019, Article ID 1518246, 12 page 2019.
- [30] "Case western reserve university bearing data center," 2019, <https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website>.
- [31] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Computational Statistics & Data Analysis*, vol. 143, Article ID 106839, 2020.
- [32] A. Kumar, C. Gandhi, X. Liu et al., "A novel health indicator developed using filter-based feature selection algorithm for

- the identification of rotor defects,” *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 2020.
- [33] S. Fong, R. P. Biuk-Aghai, and R. C. Millham, “Swarm search methods in Weka for data mining,” in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pp. 122–127, Singapore, 2018.
 - [34] M. Dorigo, V. Maniezzo, and A. Coloni, “Ant system: optimization by a colony of cooperating agents,” *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 26, no. 1, pp. 29–41, 1996.
 - [35] M. Kang, J. Kim, and J.-M. Kim, “Reliable fault diagnosis for incipient low-speed bearings using fault feature analysis based on a binary bat algorithm,” *Information Sciences*, vol. 294, pp. 423–438, 2015.
 - [36] D. T. Pham and M. Castellani, “The Bees algorithm: modelling foraging behaviour to solve continuous optimization problems,” *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 223, no. 12, pp. 2919–2938, 2009.
 - [37] X.-S. Yang and S. Deb, “Cuckoo search via Lévy flights,” in *Proceedings of the 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC)*, IEEE, Coimbatore, India, pp. 210–214, 2009.
 - [38] S. Deb, S. Fong, and Z. Tian, “Elephant search algorithm for optimization problems,” in *Proceedings of the 2015 10th International Conference on Digital Information Management (ICDIM)*, IEEE, Jeju, Republic of Korea, pp. 249–255, 2015.
 - [39] X.-S. Yang and X. He, “Firefly algorithm: recent advances and applications,” 2013, <https://arxiv.org/abs/1308.3898>.
 - [40] X.-S. Yang, “Flower pollination algorithm for global optimization,” in *Proceedings of the 2012 International Conference on Unconventional Computing and Natural Computation*, pp. 240–249, Orléans, France, September 2012.
 - [41] R. Tang, S. Fong, X.-S. Yang, and S. Deb, “Integrating nature-inspired optimization algorithms to K-means clustering,” in *Proceedings of the 7th International Conference on Digital Information Management (ICDIM 2012)*, pp. 165–172, Macau, China, 2012.
 - [42] G.-G. Wang, X.-Z. Gao, K. Zenger, and L. d. S. Coelho, “A novel metaheuristic algorithm inspired by Rhino Herd behavior,” in *Proceedings of the 9th EUROSIM Congress on Modelling and Simulation, EUROSIM 2016, The 57th SIMS Conference on Simulation and Modelling SIMS 2016*, vol. 142, pp. 1026–1033, Oulu, Finland, 2018.
 - [43] E. Frank and I. H. Witten, “Generating accurate rule sets without global optimization,” 1998.
 - [44] W. W. Cohen, “Fast effective rule induction,” in *Proceedings of the 1995 12th International Conference on Machine Learning*, pp. 115–123, Shenzhen, China, 1995.
 - [45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
 - [46] G. I. Webb, “MultiBoosting: a technique for combining boosting and wagging,” *Machine Learning*, vol. 40, no. 2, pp. 159–196, 2000.