Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 18, Iss. 2, pp. 16-26, 2020

USING THE KULLBACK-LEIBLER DIVERGENCE AND KOLMOGOROV-SMIRNOV TEST TO SELECT INPUT SIZES TO THE FAULT DIAGNOSIS PROBLEM BASED ON A CNN MODEL

Rodrigo P. Monteiro[1], **Carmelo J. A. Bastos-Filho**[2]
Federal University of Pernambuco[1], Brazil, University of Pernambuco[2], Brazil
rodrigo.paula@ufpe.br, carmelofilho@ieee.org

Mariela Cerrada[3], **Diego R. Cabrera**[3], **René V. Sánchez**[3]
Universidad Politécnica Salesiana[3], Ecuador
{mcerrada, dcabrera, rsanchezl}@ups.edu.ec

Abstract – Choosing a suitable size for signal representations, *e.g.*, frequency spectra, in a given machine learning problem is not a trivial task. It may strongly affect the performance of the trained models. Many solutions have been proposed to solve this problem. Most of them rely on designing an optimized input or selecting the most suitable input according to an exhaustive search. In this work, we used the Kullback-Leibler Divergence and the Kolmogorov-Smirnov Test to measure the dissimilarity among signal representations belonging to equal and different classes, *i.e.*, we measured the intraclass and interclass dissimilarities. Moreover, we analyzed how this information relates to the classifier performance. The results suggested that both the interclass and intraclass dissimilarities were related to the model accuracy since they indicate how easy a model can learn discriminative information from the input data. The highest ratios between the average interclass and intraclass dissimilarities were related to the most accurate classifiers. We can use this information to select a suitable input size to train the classification model. The approach was tested on two data sets related to the fault diagnosis of reciprocating compressors.
Keywords – Deep learning, Kullback-Leibler Divergence, Kolmogorov-Smirnov Test, Input Size Selection.

1. INTRODUCTION

Deep learning techniques have been widely used in signal processing. They are applied on several domains, *e.g.* speech and audio recognition [1], fault detection and diagnosis [2, 3], feature extraction on images and videos [3, 4]. Those techniques learn to represent the input information at different complexity levels along with their intermediate layers. In other words, the first layers of the network learn local and simpler patterns from the input information, while the following layers use those patterns to build more complex concepts. Such a hierarchical characterization of the input information may improve the description of objects for a given task, *e.g.* objects classification or detection [5].

The deep learning-based models may have their performance enhanced or worsened according to the kind of information they receive, *e.g.* time- or frequency-domain signals, and one- or n-dimensional inputs. Choosing the most suitable input representation for a given problem is not a trivial task. Moreover, this issue also affects other approaches than the deep learning-based ones, such as traditional machine learning, statistics, among others [6, 7].

Many solutions have been developed, aiming to solve this problem. A large group of solutions focuses on input design. One example is presented by Khandelwal *et al.* [8]. They proposed a performance index to measure the optimality of a given input for fault diagnosis. Their proposal is based on the analysis of LTI systems. Wang-Jian [9] designed input signals for multi UAVs formation anomaly detection. Furthermore, Heirung and Mesbah [10] provided an overview of the state-of-the-art methods for input design regarding the active fault diagnosis. They also discussed the primary considerations in the formulation and solution of input design problems.

Another critical group of solutions uses the input selection. It contains a wide range of techniques, which depend on the nature of the data. Most of the works found in the literature that chooses a subset of variables that belong to a more extensive set are strictly related to feature selection. One example is the work of van de Laar and Heskes [11], which performed the input selection based on an ensemble of neural networks. Regarding time series, Sorjamaa *et al.* [12] used input selection criteria such as k-Nearest Neighbors, Mutual Information, and Nonparametric Noise Estimation (NNE) in the long-term prediction of electricity loads.

In this work, we used the Kullback-Leibler (KL) Divergence and the Kolmogorov-Smirnov (KS) Test as metrics to select a suitable input size for the fault diagnosis problem. Those inputs are frequency spectra and spectrograms. To be more specific, we used the ratio between the average interclass and intraclass metric values of those inputs to infer how accurate the trained classifier would perform. This was a comparative analysis, *i.e.*, we compared ratios obtained from different input sizes and inferred which size was the most appropriate to train an accurate classification model. This approach was tested on two data sets related to the fault diagnosis of reciprocating compressors.

According to the results, the highest values of this ratio were related to the most accurate models, while values close to 1 led to poor classification performances. This fact has an explanation. If the ratio is close to one, the interclass and intraclass

dissimilarities are close to each other. So, we expect more difficulty to train accurate models. On the other hand, the higher the ratio is, the larger is the interclass dissimilarity concerning the intraclass one. So, we expect more ease to train accurate classifiers. Those results suggested that we can use the proposed approach to select suitable input sizes for the classification problem. The main advantage of this approach is not having to train classifiers for each input size to select the most suitable one. In this way, we save time and computational resources.

The remaining of the paper is organized as follows: Section II presents the Theoretical Background, Section III describes the Methodology, Section IV shows the results and discussions, and Section V gives our Conclusions.

2. THEORETICAL BACKGROUND

In this section, we present the Kolmogorov-Smirnov Test and Kullback-Leibler Divergence. Those techniques are commonly used to compare probability distributions. In this work, we use both techniques to measure the dissimilarity of signals represented by frequency spectra and spectrograms.

2.1 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov Test [13, 14] is a nonparametric statistical test that compares continuous, one-dimensional probability distributions. One can use the KS Test to investigate whether two samples come or not from statistically identical distributions (two-sample KS), or whether a sample comes or not from a reference probability distribution (one-sample KS). According to the null hypothesis, the compared samples come from statistically identical probability distributions. On the other hand, the rejection of the null hypothesis means that those samples come from different distributions.

Regarding the two-sample KS, the statistical test is the maximum of the absolute value of the difference between two empirical distribution functions, as presented by Equation (1).

$$D_{n_1,n_2} = \max_t |F_{1,n_1}(t) - F_{2,n_2}(t)| \tag{1}$$

in which F_1 and F_2 are empirical distribution functions of the first and second samples, respectively. n_1 and n_1 are the sizes of the first and second samples, respectively.

For large samples, the null hypothesis is rejected at a given significancy level α, *e.g.* 5%, if Equations (2) and (3) are true.

$$D_{n_1,n_2} > c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}. \tag{2}$$

$$c(\alpha) = \sqrt{-\frac{1}{2}\ln\alpha} \tag{3}$$

One example of work relating the KS Test to fault detection and signal processing was developed by Stepančič *et al.* [15]. They used the test to detect faults on fuel cell systems based on impedance data.

2.2 Kullback-Leibler Divergence

Also called relative entropy, the Kullback-Leibler Divergence [16, 17] measures how a probability distribution is different from a second or a reference probability distribution. Such difference, also referred to as divergence, is calculated in terms of a measure of information, *e.g.* entropy. It is a measure of surprise, with diverse applications, such as applied statistics and machine learning. For for two distributions P_1 and P_2 of a continuous random variable x, the Kullback-Leibler Divergence is defined by the Equation (4).

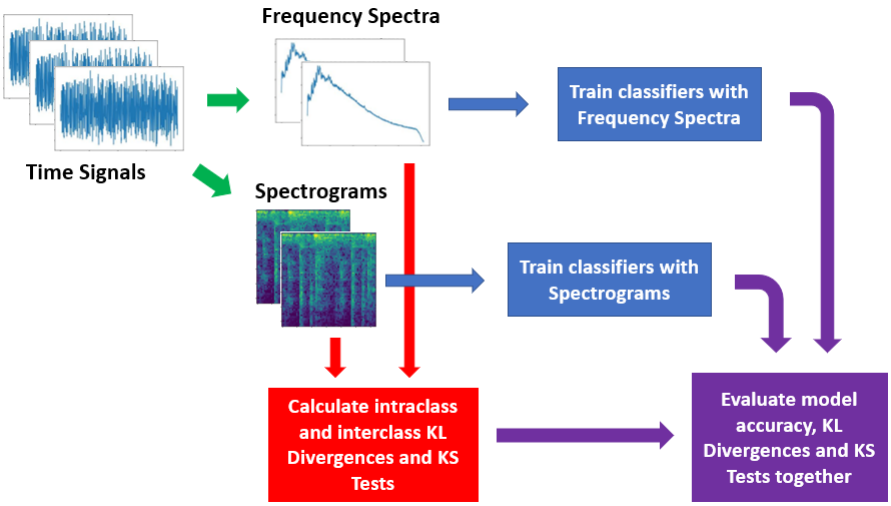
$$D_{KL}(P_1(x)||P_2(x)) = \int_{-\infty}^{\infty} p_1(x)\log\left(\frac{p_1(x)}{p_2(x)}\right)dx \tag{4}$$

in which p_1 and p_2 denote the probability densities of P_1 and P_2, respectively.

One example of work that relates the relative entropy to fault detection and signal processing was developed by Wang *et al.* [18]. They used the Kullback-Leibler divergence to detect anomalies on probability density functions. In this study, the anomalies were human behind walls.

3. Methodology

We developed this work according to the steps presented in Figure 1. At first, we preprocessed the data consisting of time-domain signals. In this step, we carried out the required transformations and obtained the frequency spectra and spectrograms, which were used to train the classification models. Also, we applied the KS Test and KL Divergence on the frequency spectra and spectrograms. Thus, we measured the dissimilarity among signals that belong to equal and different classes. Then, we analyzed those measures together with the accuracies of the classifiers. The whole process is better described in the following subsections.



LNLM.PNG

Figure 1: Process developed in this work.

3.1 Data set

The data set consists of vibration signals. Those signals were collected by accelerometers placed on a two-stage reciprocating compressor available at the Universidad Politécnica Salesiana in Cuenca, Ecuador. We regarded two experimental scenarios. In the first one, the signals are divided into four classes, as presented in Table 1, which are related to bearing faults. In the second scenario, the signals are divided into thirteen classes, which are related to bearing and valve faults (multiple faults) presented in Table 2. In both scenarios, the classes correspond to normal and faulty operation conditions. Each class contains fifteen time-domain signals, each one being 10 seconds long and collected at a 50 kHz sampling rate. The frequency range of those signals was up to 15 kHz. The complete description of the experimental setup and signal acquisition process is seen in [19, 20].

Table 1: Classes of signals regarding bearing faults.

Bearing Faults	
Fault Code	**Fault Type**
P1	No fault
P2	Inner race crack
P3	Roller element crack
P4	Outer race crack

3.2 Formatting the inputs

The time-domain signals were divided into smaller sections, each one being 0.1 second long. This resulted in 1,500 vibration signals for each class. In this work, we used two kinds of signal representations: the frequency spectrum and the spectrogram. The first one is a frequency-domain representation and allows the signal analysis in terms of its frequency components. On the other hand, the second one is a time-frequency-representation and allows the visualization of the signal frequency components over time.

The frequency spectra were obtained via the Fast Fourier Transform (FFT) [21]. We assessed the influence of different frequency resolutions on model accuracy. In this way, we tested frequency spectra with 8, 16, 32, 64, 128, 256, 512, and 1024 frequency bins. Those bins were obtained by cropping and zero-padding the time-domain signals, depending on the frequency spectrum size concerning the input signals.

On the other hand, the spectrograms were obtained via the Short-Time Fourier Transform (STFT) [21]. They carry more information concerning the frequency spectra since they show how the frequency components vary over time on a given signal. However, these representations increase the computational burden of the classification process regarding the scenario in which the inputs are frequency spectra. In this new scenario we tested 8 different frequency resolutions: 8, 16, 24, 32, 40, 48, 56 and 64. This way, the shapes of the resulting spectrograms were: 8 x 357, 16 x 178, 24 x 118, 32 x 89, 40 x 71, 48 x 59, 56 x 50 and 64 x 44. The window overlap was 12.5%.

3.3 Input dissimilarity with the Kullback-Leibler Divergence and Kolmogorov-Smirnov Test

For each data set, we calculated the KL Divergence and the KS Test between pairs of frequency spectra, and also between pairs of spectrograms. They were calculated by using Equations (1) and (4), which were presented in Section 2. Those pairs

Table 2: Classes of signals regarding multiple faults.

Multi Faults	
Fault Code	**Fault Type**
P1	No fault
P2	Bearing inner race crack/ Valve seat wear
P3	Bearing inner race crack/ Corrosion of the valve plate
P4	Bearing inner race crack/ Fracture of the valve plate
P5	Bearing inner race crack/ Spring break
P6	Bearing roller element crack/ Valve seat wear
P7	Bearing roller element crack/ Corrosion of the valve plate
P8	Bearing roller element crack/ Fracture of the valve plate
P9	Bearing roller element crack/ Spring break
P10	Bearing outer race crack/ Valve seat wear
P11	Bearing outer race crack/ Corrosion of the valve plate
P12	Bearing outer race crack/ Fracture of the valve plate
P13	Bearing outer race crack/ Spring break

included data belonging to equal and different classes, but with the same dimension. In this way, we obtained intraclass and interclass dissimilarity measures for each input shape. Those dissimilarity measures can explain how easy a classification model learn discriminative information from data. High interclass and low intraclass dissimilarities are desirable to train accurate classification models.

In this work, we used the ratio between the average interclass and intraclass dissimilarity measures to infer how accurate a model would perform. This was a comparative analysis. In other words, we compared the ratios obtained from different input shapes and inferred which input shape was the most appropriate to perform the fault diagnosis. Our hypothesis is that input data presenting a higher ratio lead to a more accurate classification model. In this sense, if this ratio is close to one, the interclass and intraclass dissimilarities are close to each other. So, we expect more difficulty to train accurate models. On the other hand, the higher the ratio is, the larger is the interclass dissimilarity concerning the intraclass one. So, we expect more ease to train accurate models.

Those ratios were calculated for both KL Divergence and KS Test dissimilarity measures.

3.4 Model

We used the convolutional neural network (CNN) [22] to perform the classification process. We adopted an architecture with one convolutional layer, one max pooling layer, one flattening layer, one densely connected layer, and one output layer. The layers are better described in Table 3. Since the objective of this work is to evaluate how the type of input representation influences on the classifier performance, the choice of the model parameters is not critical for the experiments.

3.5 Experiments

As explained in Subsection 3.2, we divided the dataset into 4 classes regarding the first fault scenario and 13 classes regarding the second one. Each class in both scenarios had 1,500 signals. Those signals were divided into 10 subsets, which were combined randomly to form the training and test sets. 80% (1,200 signals or 8 subsets) of the signals were allocated to the training set, while the remaining 20% (300 signals or 2 subsets) were allocated for the testing set. Each model was trained and tested with a different combination of subsets.

We aim to assess the influence of different input shapes on the model performance, regarding faults scenarios with different complexities. In this analysis, we used the Kullback-Leibler Diverenge and the Kolmogorov-Smirnov test, which were calculated between pairs of frequency spectra or spectrograms. We trained 15 models for each fault and input shape scenario, aiming to achieve results with statistical relevance. The number of training epochs was 50. The configuration of the computer used to

Table 3: CNN configuration.

Layers
Convolutional layer with 8 filters (3 x 3 for spectrograms, 3 x 1 for frequency spectra)
Max pooling layer (2 x 2 for spectrograms, 2 x 1 for frequency spectra)
Flattening layer
Densely connected layer (32 neurons)
Output layer (4 neurons - Bearing Faults, 13 neurons - Multi Faults)

train the models was: OS Windows 10 Home, 64 bits, Memory (RAM) 15.9 GB, Processor Intel® Corte™ i7-6500 CPU @ 2.50 GHz x 2, AMD Radeon™ T5 M330 (No CUDA support). All the scripts were written in Python [23] 3.7, on the Jetbrains PyCharm [24] Community Edition 2019.2.

4 Results and Discussions

4.1 Frequency Spectra

First, we analyzed the accuracy of those models trained to classify bearing faults. As seen in Table 1, there are four classes in this scenario. Table 4 lists the average accuracy of the classifiers trained with frequency spectra. The results were arranged according to the input size, and they are related only to test samples.

Table 4: Average accuracy for different frequency spectrum sizes for bearing faults.

Input Size	**Accuracy**
8	0.4613
16	0.6009
32	0.7142
64	0.8204
128	0.8987
256	0.9542
512	0.9852
1024	0.9983

As seen in Table 4, the classifiers became more accurate as the input size increased, *e.g.* the accuracy increased by 116.41% (from 0.4613 to 0.9983) when the input size varied from 8 to 1024. One reason may be the higher amount of information available when the input size increases. A second reason may be the quality of the information that each input size carries. In other words, the signals are better represented in the frequency-domain when the frequency spectrum presents a higher resolution. In this way, the frequency spectra may carry more discriminative information.

We used two dissimilarity metrics to assess the quality of the information provided by different input sizes, and also to better understand the performance of the classification models. Those metrics were the Kullback-Leibler Divergence and the Kolmogorov-Smirnov Test. We have calculated the ratio (R) between the average interclass and intraclass dissimilarities, regarding both metrics.

Tables 5 and 6 list the average KL Divergence and KS Test between pairs of classes, respectively. Those results concern the frequency spectra with size 1024 and bearing faults. To obtain the ratio R, we first calculate the average interclass and intraclass dissimilarities. Regarding the average intraclass KL Divergence, we calculate the average value of the main diagonal elements in Table 5. Those elements represent the average KL Divergence between frequency spectra belonging to the same classes, *e.g.*, P1 and P1, P2 and P2, etc. On the other hand, we obtain the interclass KL Divergence by calculating the average value of the remaining table elements, that is, the dissimilarity between different classes, *e.g.*, P1 and P2, P2 and P3, etc. In this way, we achieved a value of 0.0153 for the average intraclass and 0.0223 for the average interclass KL Divergences. Thus, the ratio R is 1.4575. Following the same process for the KS Test and using the results presented in Table 6, we obtained a ratio equal to 1.4964.

By repeating this process for all input sizes, we obtain the results presented in Figures 2 and 3. The green bars refer to the accuracy values listed in Table 4. The blue bars, on the other hand, refer to the ratio R of the KL Divergence (Figure 2) and KS Test (Figure 3) for different sizes. We observe in those figures that the accuracy increased as the ratio R increased. It occurred to both KL Divergence and KS Test, suggesting that the ratios between the average interclass and intraclass values of those two metrics may relate to the ease or difficulty of training the accurate models.

Table 5: Average KL Divergence regarding each pair of classes individually, for the input size equal to 1024 in the frequency spectra scenario

Classes	P1	P2	P3	P4
P1	0.0156	0.0248	0.0171	0.0195
P2	0.0241	0.0178	0.0262	0.0298
P3	0.0174	0.0271	0.0141	0.0158
P4	0.0197	0.0308	0.0157	0.0137

Table 6: Average KS Test regarding each pair of classes individually, for the input size equal to 1024 in the frequency spectra scenario.

Classes	P1	P2	P3	P4
P1	0.1205	0.1958	0.1382	0.1299
P2	0.1958	0.1032	0.2064	0.2094
P3	0.1382	0.2064	0.1078	0.1115
P4	0.1299	0.2094	0.1115	0.1100

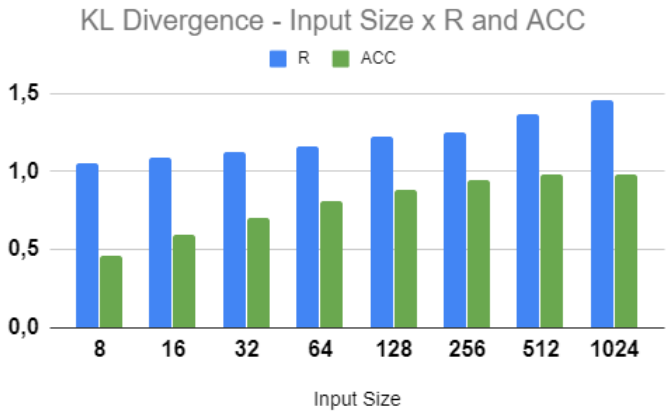


Figure 2: KL Divergence ratio and classifier accuracy for different input shapes regarding bearing faults.

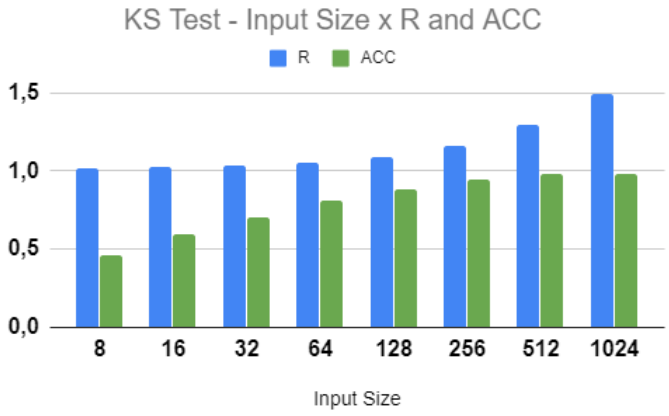


Figure 3: KS Test ratio and classifier accuracy for different input shapes regarding bearing faults.

We can also perform this analysis on the multiple faults data set. This scenario is more complex than the first one since it presents 13 fault modalities, as seen in Table 2. The inputs are also frequency spectra and their sizes vary from 8 to 16. Table 7 lists the average accuracy of the classifiers trained with multiple faults data. The results were arranged according to the input size, and they are related only to test samples.

Table 7 shows a trend similar to the one presented in Table 4. In other words, the classifiers became more accurate as the input size increased. The accuracy increased by 475.16% (from 0.1236 to 0.7109) when the input size varied from 8 to 1024. On the other hand, the best accuracy value was smaller than the one achieved by those models trained with bearing faults data. We repeated the analysis performed with bearing faults to better understand the performance of classifiers on this new data set. The results are presented in Figures 4 and 5. The green bars refer to the accuracy values listed in Table 7. The blue bars, on the other

Table 7: Average accuracy for different frequency spectrum sizes for multiple faults.

Input Size	Accuracy
8	0.1236
16	0.1537
32	0.2237
64	0.3234
128	0.4373
256	0.5724
512	0.6948
1024	0.7109

hand, refer to the ratio R of the KL Divergence (Figure 4) and KS Test (Figure 5) for different sizes.

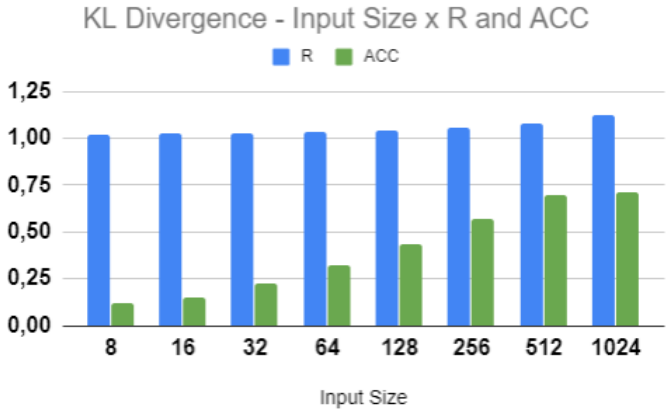


Figure 4: KL Divergence ratio and classifier accuracy for different input shapes regarding multiple faults.

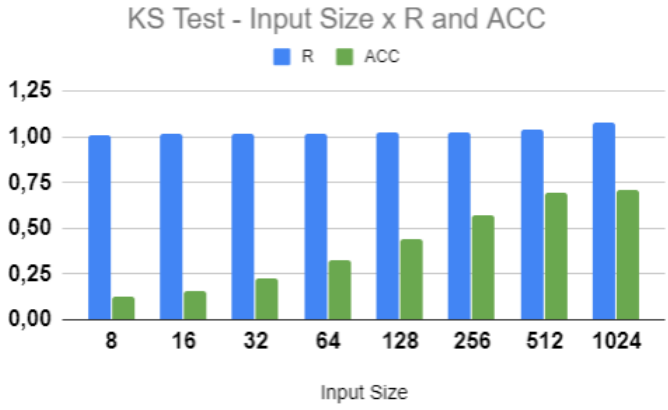


Figure 5: KS Test ratio and classifier accuracy for different input shapes regarding multiple faults.

We observe in those figures that the accuracy increased as the ratio R increased, similar to what occurred to the bearing faults. It also happened to both KL Divergence and KS Test, reinforcing the hypothesis that those ratios between the average interclass and intraclass metric values may relate to the ease or difficulty of training the accurate models. Besides, the ratios in this scenario are lower than the ones of bearing faults, which may explain the lower accuracy values for multiple faults. Indeed, when the ratio is close to 1, the average dissimilarity of spectra belonging to different classes is close to the one of same classes. It makes the learning process of discriminative patterns more difficult, demanding the use of more powerful classification models to achieve a satisfactory performance level.

4.2 Spectrograms

The first analysis in this subsection regards the bearing faults, which are divided into four classes. This time the input data are spectrograms. Unlike the frequency spectra, their information is organized in the form of two-dimensional matrices. As

previously mentioned, they show how the frequency components of the signal vary over time. Table 8 lists the average accuracy of classifiers trained with bearing faults spectrograms. The results were arranged according to the input size, and they are related only to test samples.

Table 8: Average accuracy for different spectrogram sizes for bearing faults.

Input Size	Accuracy
8 x 357	0.9980
16 x 178	0.9994
24 x 118	1
32 x 89	1
40 x 71	1
48 x 59	1
56 x 50	1
64 x 44	1

The trend seen in Table 8 is similar to the one observed in Tables 4 and 7. In other words, the accuracy improved as the frequency resolution increased. However, in Table 8 the classifiers achieved a 100% accuracy with lower frequency resolutions. It probably occurred because of the complementary time information provided by spectrograms. We also analyzed the information present on the spectrograms by using the ratio between the average interclass and intraclass dissimilarity measures, *i.e.*, the KL Divergence, and KS Test. Those results are seen in Figures 6 and 7.

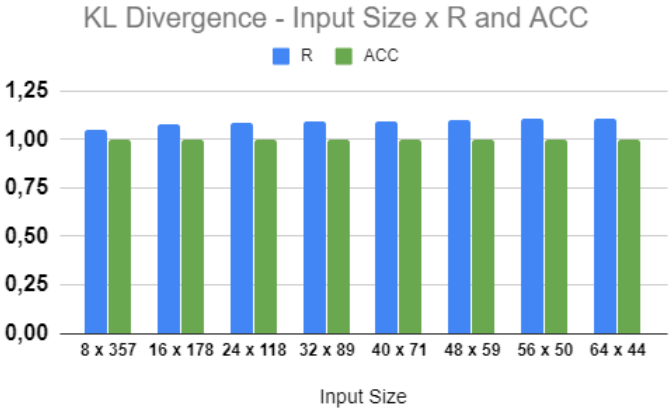


Figure 6: KL Divergence ratio and classifier accuracy for different spectrogram shapes regarding bearing faults.

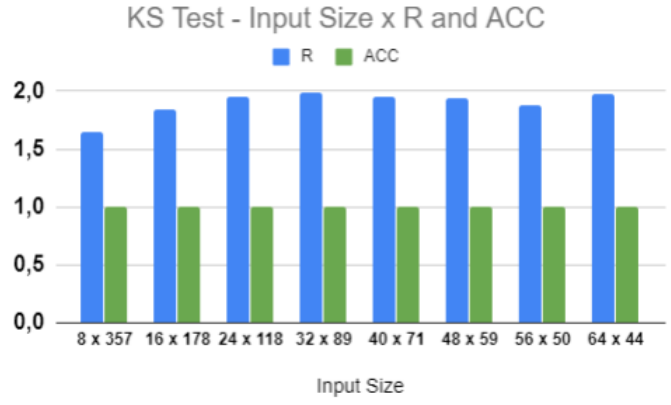


Figure 7: KS Test ratio and classifier accuracy for different spectrogram shapes regarding bearing faults.

Table 8 and Figures 6 and 7 show that the model accuracies reached their maximum value for most input shapes. The exceptions were the input shapes 8 x 357 and 16 x 178, which presented accuracy values equal to 0.9980 and 0.9994. Still, these two shapes were the ones related to the lowest R values. Regarding the remaining shapes, the analysis became unnecessary, since a further increase of R above a given value, *e.g.*, 1.75 for the KS Test, did not result in accuracy variations.

We also performed this analysis on the multiple faults data set, *i.e.*, the one with 13 fault modalities. The inputs are also spectrograms whose input shapes varied from 8 x 357 to 64 x 44. Table 9 lists the average accuracy of the classifiers trained with those spectrograms. The results were arranged according to the input size, and they are related only to test samples.

Table 9: Average accuracy for different spectrogram sizes for multiple faults.

Input Size	Accuracy
8 x 357	0.6171
16 x 178	0.7744
24 x 118	0.8095
32 x 89	0.8567
40 x 71	0.8660
48 x 59	0.8837
56 x 50	0.8916
64 x 44	0.9127

In this scenario, the accuracy values improved as the frequency resolution increased, just as in the experiments of the past. Figures 8 and 9 show how the ratio between the interclass and intraclass dissimilarity measures varies over time. We also observe that both the ratio R and the classifier accuracy increased as the input shape varied from 8 x 357 to 64 x 44. However, the ratio increased at a rate much smaller than the accuracy one. It probably occurred due to the higher number of classes that the multiple faults data set presents.

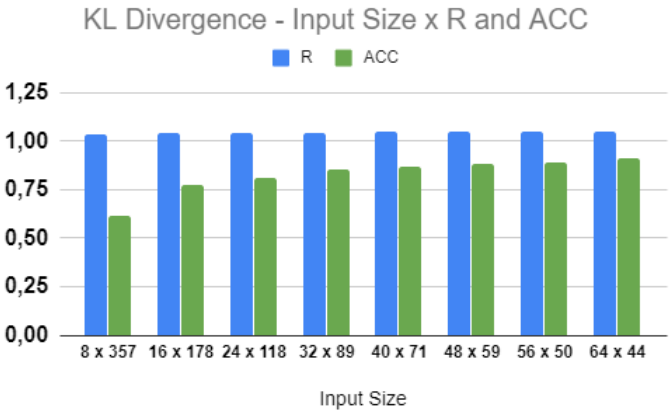


Figure 8: KL Divergence ratio and classifier accuracy for different spectrogram shapes regarding multiple faults.

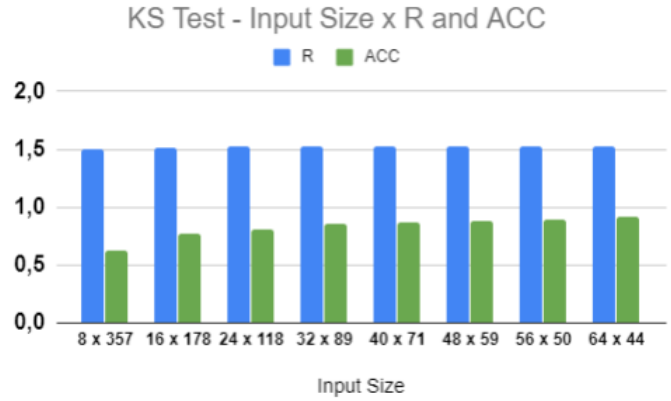


Figure 9: KS Test ratio and classifier accuracy for different spectrogram shapes regarding multiple faults.

The results obtained by those experiments led us to some considerations. The first one regards the performance of the classifiers on each data set. The models trained to classify bearing faults were the ones that presented the highest accuracy values. This result corroborates our hypothesis since the frequency spectra and spectrograms related to bearing faults presented the highest ratios between the interclass and intraclass dissimilarity measures.

The second consideration concerns the ratio and classifier performance for different input sizes. Although the observed behaviors were more accentuated on frequency spectra than on spectrograms, there was a global trend in which the accuracy tended to improve as the ratio between the interclass and intraclass dissimilarity measures increased. We observed such a trend on both dissimilarity metrics analyzed, *i.e.*, the Kullback-Leibler Divergence, and the Kolmogorov-Smirnov Test. This was a piece of relevant information since it can help to save time when we need to choose the most suitable input format for a given classification problem. Many works in the literature deal with this issue by training models on different input shapes and choosing the one that led to the highest accuracy values. Such a process can be quite time consuming, and also demand high computational resources. On the other hand, the ratio between the interclass and intraclass divergence measures could help in this choice without training the classifiers or performing a grid search. We could make this choice by comparing the proposed ratios obtained from data with different input shapes.

One example is presented in Figures 2 and 3. The ratio R obtained from the input size 8 was very close to 1. It means that the intraclass dissimilarities are almost equal to the interclass ones. It suggests that the trained model tends to face more difficulties to learn discriminative patterns due to the poor data separability. Then, the accuracy levels of the trained model tend to be unsatisfactory. On the other, the input size of 1024 led to a ratio close to 1.5. It means that the interclass dissimilarities are higher concerning the intraclass ones, suggesting that the model learning process will be easier and more effective. The same logic can be applied to the remaining scenarios.

5. CONCLUSIONS

This work proposed to use dissimilarity metrics to assess the information contained on different signal representations. We also analyzed how those metrics relate to the classifier accuracy. The signal representations were spectrograms and frequency spectra, and the metrics were the Kullback-Leibler Divergence and the Kolmogorov-Smirnov Test. We used the ratio between the average interclass and intraclass dissimilarity measures to choose the input size that possibly lead to the most accurate classifiers. This was a comparative analysis. In other words, we needed to calculate the ratio for different input sizes. Then, we compared the resulting values to choose the most suitable size.

We performed the analysis on two data sets. They consisted of vibration signals collected by an accelerometer placed on a reciprocating compressor. The first data set had four classes and regarded bearing faults. The second one had 13 classes and regarded multiple faults, *i.e.* bearing and valve faults. The results achieved on both data sets suggested that the ratios calculated were related to the accuracy of the trained classifiers. It occurred for the KL Divergence and KS test, and also for inputs represented by frequency spectra and spectrograms. One example was the scenario with bearing faults data represented by frequency spectra. The accuracy increased from 0.4613 to 0.9983 when the input size varied from 8 to 1024. Also, the ratio varied from 1 to about 1.5 for the KS Test, meaning an improvement in data separability, which could explain the accuracy raise.

The ratio between the interclass and intraclass dissimilarity measures can aid the choice of the most suitable input shape for a given problem. We make this choice without training the classifiers or performing a grid search, but only by comparing the proposed ratio obtained from data with different input shapes. In this way, we save time and computational resources.

Regarding future works, we intend to extend this analysis to time-domain signals and a more significant number of data sets, also optimizing other parameters of the deep learning-based model. The new data sets are related to different nature signals, *e.g.* acoustic and electrical signals, and different nature faults, *e.g.* gearboxes and other mechanical devices. This way, we aim to verify the generalization capability of this kind of analysis.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and Universidad Politécnica Salesiana through the research group GIDTEC.

REFERENCES

[1] D. Yu and L. Deng. "Deep learning and its applications to signal and information processing". *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, 2010.

[2] R. P. Monteiro, M. Cerrada, D. R. Cabrera, R. V. Sánchez and C. J. Bastos-Filho. "Using a support vector machine based decision stage to improve the fault diagnosis on gearboxes". *Computational intelligence and neuroscience*, vol. 2019, 2019.

[3] R. P. Monteiro and C. J. Bastos-Filho. "Feature Extraction Using Convolutional Neural Networks for Anomaly Detection." In *CBIC 2019: XIV Congresso Brasileiro de Inteligência Computacional*, Belém, Brazil, Nov. 2009.

[4] M. Z. Uddin, M. M. Hassan, A. Almogren, M. Zuair, G. Fortino and J. Torresen. "A facial expression recognition system using robust face features from depth videos and deep learning". *Computers & Electrical Engineering*, vol. 63, pp. 114–125, 2017.

[5] I. Goodfellow, Y. Bengio and A. Courville. *Deep learning*. MIT press, 2016.

[6] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[7] G. James, D. Witten, T. Hastie and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[8] D. Khandelwal, S. Weiland and A. Khalate. "Robust fault diagnosis by optimal input design for self-sensing systems". *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 1031–1036, 2017.

[9] H. Wang-jian. "Optimal input design for multi UAVs formation anomaly detection". *ISA transactions*, vol. 91, pp. 157–165, 2019.

[10] T. A. N. Heirung and A. Mesbah. "Input design for active fault diagnosis". *Annual Reviews in Control*, 2019.

[11] P. Van De Laar and T. Heskes. "Input selection based on an ensemble". *Neurocomputing*, vol. 34, no. 1-4, pp. 227–238, 2000.

[12] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji and A. Lendasse. "Methodology for long-term prediction of time series". *Neurocomputing*, vol. 70, no. 16-18, pp. 2861–2869, 2007.

[13] M. Neuhauser. *Nonparametric statistical tests: A computational approach*. Chapman and Hall/CRC, 2011.

[14] D. E. Knuth. *Art of computer programming, volume 2: Seminumerical algorithms*. Addison-Wesley Professional, 2014.

[15] M. Stepančič, D. Juričić and P. Boškoski. "Fault detection of fuel cell systems based on statistical assessment of impedance data". *Energy Conversion and Management*, vol. 195, pp. 76–85, 2019.

[16] S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[17] S. Kullback and R. A. Leibler. "On information and sufficiency". *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[18] W. Wang, B. Zhang, D. Wang, Y. Jiang, S. Qin and L. Xue. "Anomaly detection based on probability density function with Kullback–Leibler divergence". *Signal Processing*, vol. 126, pp. 12–17, 2016.

[19] D. Cabrera, A. Guamán, S. Zhang, M. Cerrada, R.-V. Sánchez, J. Cevallos, J. Long and C. Li. "Bayesian approach and time series dimensionality reduction to LSTM-based model-building for fault diagnosis of a reciprocating compressor". *Neurocomputing*, vol. 380, pp. 51–66, 2020.

[20] D. Cabrera, F. Sancho, J. Long, R.-V. Sánchez, S. Zhang, M. Cerrada and C. Li. "Generative adversarial networks selection approach for extremely imbalanced fault diagnosis of reciprocating machinery". *IEEE Access*, vol. 7, pp. 70643–70653, 2019.

[21] K. R. Rao, D. N. Kim and J. J. Hwang. *Fast Fourier transform-algorithms and applications*. Springer Science & Business Media, 2011.

[22] W. Rawat and Z. Wang. "Deep convolutional neural networks for image classification: A comprehensive review". *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[23] G. VanRossum and F. L. Drake. *The python language reference*. Python Software Foundation Amsterdam, Netherlands, 2010.

[24] Q. N. Islam. *Mastering PyCharm*. Packt Publishing Ltd, 2015.