CNN-DFT Based Approach Applied to Image Inspection of Railcar Component: A Comparison with Machine Learning Methods

Rafael L. Rocha[1], Cleison D. Silva[1],
Ana C. S. Gomes[2], Bruno V. Ferreira[2], Eduardo C. Carvalho[2],
Ana C. Q. Siravenha[3], Carolina C. Rosa[1]

[1] Federal University of Para, Brazil
rafael.rocha@itec.ufpa.br, {cleison, carolinarosa}@ufpa.br
[2] SENAI Innovation Institute for Mineral Technologies, Brazil
{claudia.isi, bruno.isi}@senaipa.org.br,
eduardo.isi@sesipa.org.br
[3] Vale Institute of Technology, Brazil
ana.siravenha@pq.itv.org

Abstract. The railcar component inspection is one of the most critical tasks in railway maintenance. The use of image processing, coupled with machine learning has emerged as a solution for replacing current standard methodologies. The spectral analysis gives the frequency representation of a signal and has been largely used in signal processing tasks. In this sense, this work proposes the evaluation of the use of the discrete Fourier transform (DFT) in addition to the spatial representation image of railcar component for an automatic detector of defective parts performed by convolutional neural network (CNN) classification. The most appropriate combination of images of the spatial and frequency domains is compared to the histogram of oriented gradients (HOG) feature descriptor linked to the multilayer perceptron (MLP) and support vector machine (SVM) classification, where data augmentation is investigated to improve the classification performed by all approaches. A search is made for the parameters that best fit the MLP and SVM models for comparison with the proposed approach. The results are given in measure of accuracy in addition to accuracy boxplot, and it showed encouraging results in the combination of spatial image and DFT magnitude combined with data augmentation as CNN inputs, reaching an accuracy of 96.04% and demonstrating statistically to have a significant difference between the comparative methods.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: railcar inspection, convolutional neural network, discrete Fourier transform, image classification

1. INTRODUCTION

The railcar is one of the most important assets of a railway since it can be used both in the cargo and people transportation. Given their role in this context, it is necessary to inspect key components in the operation of the railcar, in order essentially to prevent accidents [Macucci et al. 2016].

In a part of the branch companies, the component inspection task is performed by an operating technician located in an area on the railway assigned to this task. Because of technician distraction, fatigue, or stress, visual inspection is susceptible to inspection errors. Also, there are risks to the operating technician inherent in the unhealthy inspection environment [Hart et al. 2008; Park et al. 1996].

In industry, frequency analysis has been supporting image classification tasks for the inspection of certain objects, such as fabric defect detection, where fabric image in the spectral domain is classified through statistical and morphological processes [Sakhare et al. 2015].

The two-dimensional (2-D) Discrete Fourier Transform (DFT) has supported breast cancer inspection through imaging, where images features are extracted and classified by methods commonly used in machine learning [Spanhol et al. 2016; Samant and Sonar 2018]. Data augmentation is used to increase the number of training images through transformations, enabling the improvement of classification performance and the reduction of overfitting.

The purpose of this work is to investigate the frequency information in the inspection of the analyzed railcar component, the pad that plays an important role in railcar dynamics [IWnicki 2006]. As well as to evaluate the data augmentation for classification improvement. The objective is, from a convolutional neural network (CNN), to recognize patterns associated with defects in the railcar component from images of the spatial (original image) and the frequency domains, obtained from the discrete Fourier transform (DFT). The network should be able to indicate whether there are defects in the component. The possible states to be identified are (1) absent pad, (2) undamaged pad, i.e., the part is in perfect condition, and (3) damaged pad, when there is an indication of damage that can cause some inconvenience. In particular, the absence of the pad is due to the difference between the railcar designs of different manufacturers that travel these routes.

The methodology used in this work aims to investigate separately and the combination of spatial and frequency images as CNN inputs and discuss the performance improvement presented by them. The proposed approach is compared with the support vector machine (SVM) and artificial neural network (ANN) classifiers, combined with the histogram of oriented gradients (HOG) feature descriptor. The artificial expansion of images by data augmentation in both approaches is investigated. The results are evaluated by accuracy and accuracy boxplots. Besides, a statistical analysis is performed to investigate the significant difference between the results.

2. RELATED WORK

The study of Odanovic [2017] shows the component inspection in the transport of ores, in this case, a fatigue study on railway axes was carried out, and for calculating fractures up plane deformation, it came to identify how maintenance problems were decisive in the task for identifying the problem of broken components, which was simply the lack of proper maintenance for them. The study states, through heat and structural fatigue tests, that although parts appear to be worn, they are fit for use. These tests were programmed in a virtual environment. The use of the piece in question would be left to the specialists.

Haidari and Tehrani [2015] seeks to identify fatigue with the aid of cracks in train wheels, in the case of the article, it presents a train wheel in a simulated environment to identify a period for its maintenance. Each wheel would have a set of two brakes for the analysis of the temperature effects in this scenario, and for that, two tests were performed, the temperature test and the mechanical test. The thermal identification of the braking points was fundamental to the results obtained in the study, with the comparison of the vertical angular forces of the wheels it was possible to verify that the braking points are the sensitive points to the functioning of a train wheel.

Data augmentation is used to increase the training images number, as well as to increase convolutional neural network performance and to reduce model training overfitting. In [Cha et al. 2018], the data augmentation is used to increase, through horizontal flips, the images number to concrete structure cracks inspection. Vertical and horizontal mirroring is also performed, as well as random cropped to increase images training number for rail fastener classification through a convolutional neural network as show Gibert et al. [2017].

Both the one-dimensional DFT and the two-dimensional DFT are used for feature extraction and then perform image classification. Image classification of the human body posture based on a neural fuzzy network is performed by Juang and Chang [2007], the one-dimensional (1-D) DFT is applied to the vertical and horizontal histograms of the posture silhouette image, so the coefficients obtained from the 1-D DFT compose the feature vector classified by the neural fuzzy network. On the other hand, the features can be extracted from the two-dimensional (2-D) DFT and used in the classification of images, as in the texture classification through the hamming distance-based neural network [Tao et al. 2003].

In the scope of the breast cancer imaging inspection, the 2-D DFT is used as a support for the image classification task. Regarding the breast tumor tissue analysis, it is noteworthy [Spanhol et al. 2016], which focuses on tumor tissue (benign and malignant) image classification through the feature vector creation based on the extraction of characteristics by Local Phase Quantization (LPQ) of the 2-D DFT texture. The discrete cosine transform (DCT) and 2-D DFT are used in the mammogram classification for the tumor detection (benign, malignant and, normal), as shown Samant and Sonar [2018], in which DCT or DFT are applied to the mammographic images and then the gray-kevel co-occurrence matrix (GLCM) features are extracted for classification using the SVM or k-nearest neighbors (KNN), where the 2-D DFT and SVM combination achieves an accuracy of 93.89%.

In contrast to the studies cited above, which use the features extracted from the DFT to perform classification, there is a study that presents a spectral and spatial domain approach to fabric defect detection presented in [Sakhare et al. 2015]. The fast Fourier transform (FFT) and DCT are used to represent the fabric image in the spectral domain and then the classification is performed through the spatial domain statistical and morphological processes.

3. MATERIAL AND METHODS

3.1 Statement of the problem

A railway truck (Fig. 1a) is a structure underneath a railcar or locomotive to which axles (and, hence, wheels) are attached through bearings. Two trucks are fitted to each railcar or locomotive end. Typically each railcar is a 4-wheeled truck that provides support for the railcar body and is used to provide its traction and breaking [IWnicki 2006].

The analyzed railcar component, known by railway specialists as pad (Fig. 1b), is positioned between each of the side frame pedestals and the corresponding roller bearing adapter. The pad consists of a polymer inserted over the bearing adapter and plays an important role in railcar dynamics as primary suspension.

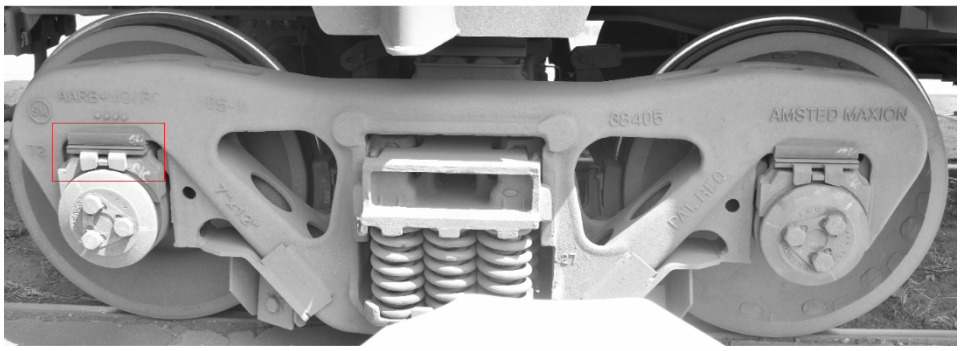
The pad image is obtained by cameras at the ore extraction site, where the railcar is slower, making it possible to capture the component image. It is worth mentioning that in this time interval, the operational technician needs to analyze several components, the pad included, hence the need to automate the component inspection by images.

3.2 Dataset

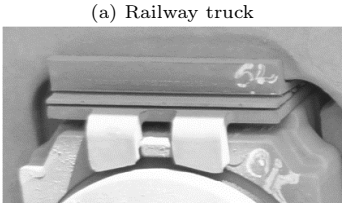
The dataset used for railcar component inspection has images describing the possible states that the pad is found in operation, namely: absent pad (Fig. 2a), undamaged pad (Fig. 2b), and damaged pad (Fig. 2c).

In particular, the pad absent comes from the design of railcar from different manufacturers, and the damage found in the pad is due to rupture or displacement from its expected position.

These three states represent the three classes to be classified in this paper, where classes 1, 2, and 3 depict the absent pad, undamaged pad, and damaged pad states.



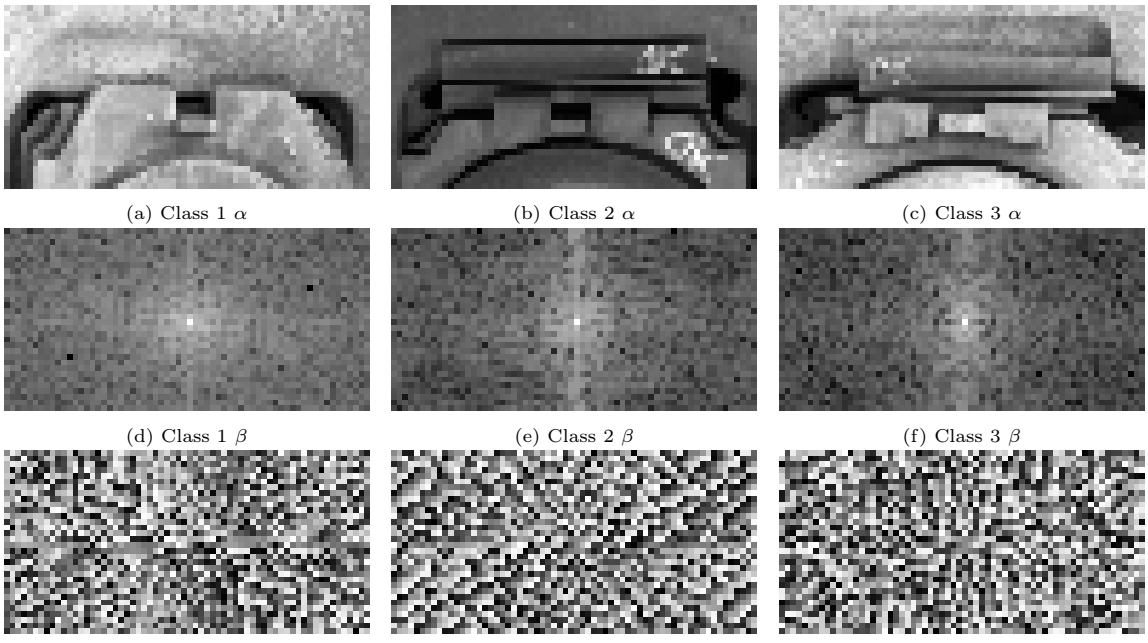
(a) Railway truck



(b) Pad

Fig. 1: View of a truck present in a railcar and the component analyzed in this paper. The red rectangle to the left of (a) highlights the investigated component, which is enlarged and further detailed in (b).

The dataset originally contains a total of 1976 spatial images of 32×64 pixels, which is divided into 651 images for class 1, 644 images for class 2, and class 3 with 681 images.



(a) Class 1 α (b) Class 2 α (c) Class 3 α

(d) Class 1 β (e) Class 2 β (f) Class 3 β

(g) Class 1 γ (h) Class 2 γ (i) Class 3 γ

Fig. 2: Input images of dimension 32×64. The images (a), (b) e (c) represent spatial images of three distinct classes; (d), (e) and (f) are frequency domain images obtained from DFT magnitude; while (g), (h) and (i) are DFT phase. Both (d)-(f) and (g)-(i) represent the same classes of (a)-(c).

3.3 Proposed approach

In this work, we propose the use of both spatial and frequency domain images for pad inspection. The spatial image is the image of the component itself, which in this work is denoted by α as shown in Figs. 2a-2c, representing classes 1, 2, and 3 respectively.

Frequency domain analysis is performed from the coefficients resulting from the DFT of the images. Considering that the quantities obtained by DFT are complex, it is essential to work with DFT in terms of the magnitude and phase quantities.

Thus, the magnitude and phase images, defined in this work by β and γ of the spatial domain image will be considered in the inspection by classifying the component pad. Figs. 2d-2f and Figs. 2g-2i demonstrates the DFT magnitude and phase of spatial images by the class of Figs. 2a-2c, respectively, which are examples taken from the dataset used in this work.

CNN's architecture of the proposed approach by this paper has six layers and is shown in Fig. 3. The input layer can have up to three images of size 32×64, which in Fig. 3 is represented by $32 \times 64 \times 3$.

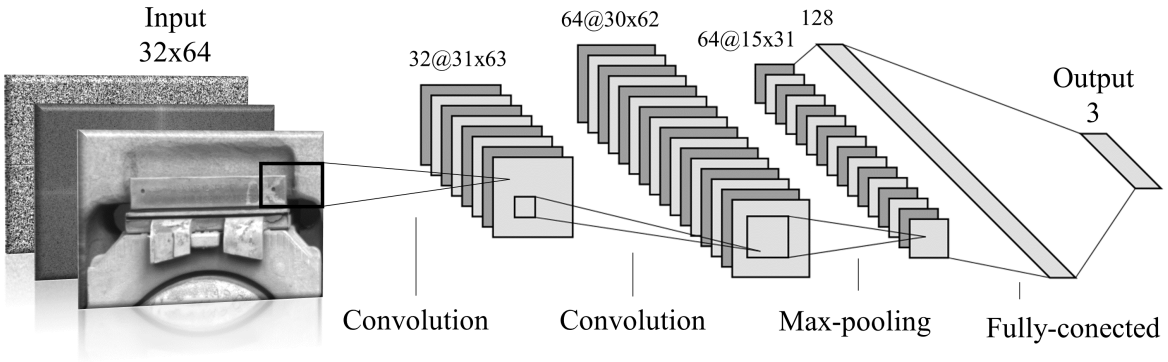


Fig. 3: The architecture of the six-layer convolutional neural network used in the experiments. The network input exemplifies the use of the component image (front) and the magnitude (middle) and phase (rear) of the discrete Fourier transform.

The first convolutional layer has 32 feature maps of size 31×63, 2×2 kernel with stride of 2 and rectified linear unit (ReLU) as the activation function. The next convolutional layer differs only in the number of feature maps and their size, which are 64 and 30×62, respectively. The max-pooling layer reduces the features previously extracted through a 2×2 kernel and stride of 2, resulting in feature maps of size 15×31.

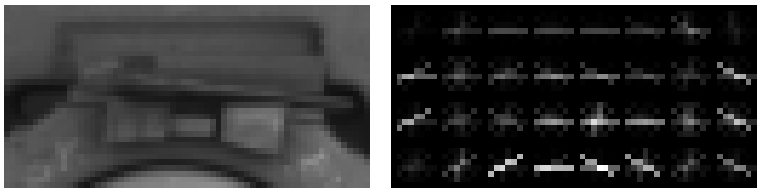
After the feature extraction process done in the convolutional and max-pooling layers, the feature maps are transformed into a feature vector that is used as input to the ANN with 128 hidden layer neurons and ReLU activation function. The feature vector has 29760 attributes, which are obtained by the number of maps by the size of each one, or $15 \times 31 \times 64$. The network output has three neurons and a softmax activation function.

3.4 Comparative methods

This subsection presents the concepts and configurations linked to the comparative methods. The feature extraction approach (3.4.1) and the data normalization mechanism (3.4.2) used to the comparative methods are presented. Also, the classification algorithms used in the comparison with the proposed approach are presented in 3.4.3 and 3.4.4.

3.4.1 *Histogram of oriented gradients.* The proposed approach, which extracts the features through the convolution and pooling process, is compared with the use of the HOG descriptor to create the feature vector evaluated by classifiers commonly employed in machine learning. The HOG is based on the evaluation of normalized local histograms of the image gradient orientations, whose idea is that the appearance and shape of the local object can be satisfactorily characterized by the distribution of local intensity gradients or edge directions without knowledge of the corresponding gradient or edge position [Dalal and Triggs 2005].

The HOG descriptor is computed initially considering the discretization of each pixel orientation in 9 histogram bins. The image window is divided into smaller regions called cells of 8×8 size, which accumulate a local histogram of gradient directions or edge orientations across the cell pixels. To normalize the contrast of responses, the local histogram measurement is accumulated in 2×2 size blocks and all cells in the block will be normalized through L2-Hys normalization, which uses L2 normalization, limiting values to 0.2 and then renormalizes by L2.



(a) Input image (b) Oriented histograms

Fig. 4: HOG features of a damaged pad. Input image (a) and the oriented histograms (b) obtained by 9 bins, cells of 8×8, and blocks of 2×2.

Fig. 4 shows the HOG features. From 9 histogram bins, cells of size 8×8 and L2-Hys normalized blocks of size 2×2 applied to the input image (Fig. 4a), and oriented histograms extracted according to these parameters are shown in Fig. 4b.

3.4.2 *Feature vector and data normalization.* The feature vector is formed by joining the HOG descriptors of the spatial α and frequency (β or γ) domains images, generating a vector of size 1512, with 756 features of α and 756 features of β or γ.

The unit used in the feature vector can significantly affect data analysis in classification algorithms such as Artificial Neural Networks, so it is necessary to normalize the feature vector data. Also, normalization helps to prevent features with values that are too large or too small [Han et al. 2011].

$$V' = \frac{V - \mu}{\sigma} \tag{1}$$

In this work, the feature vector is adjusted by z-score normalization, which normalizes the features based on the mean and standard deviation of the feature vector. The Eq. 1 demonstrates normalization z-score, where V and V' are the feature vectors before and after normalization, and μ and σ are the mean and standard deviation of the original feature vector.

3.4.3 *Artificial neural network.* The basic structure of an ANN is composed of the input, hidden and output layers, which is called multilayer perceptron (MLP). In an MLP, each neuron of a certain layer is connected to the neurons of the previous layer, and each of these neurons has a non-linear activation function.

The backpropagation learning algorithm is widely used in MLPs primarily because of its domain-independent generalization power, and alongside optimizer supports the ANN training process.

The MLP architecture used in this work has three layers: the input layer that has the feature vector with 1512 attributes, the hidden layer with 100 neurons, and the output layer formed by three neurons representing the three classes used in this work. Are investigated three activations functions in the hidden layer: Logistic, ReLU, and Tanh, which limits the values applied to neurons.

The evaluated optimizers are L-BFGS, stochastic gradient descent (SGD), and Adam. L-BFGS is a nonlinear optimizer that belongs to the family of quasi-Newton methods, a limited memory version (less data) of BGFS [Avriel 2003]. SGD is useful for using large data sets, and weights updating is performed by small random batches of data [Bishop 2006]. Adam is a variation of SGD, which uses only first-order gradients that require a small amount of memory [Kingma and Ba 2014].

3.4.4 *Support vector machine.* The SVM is a supervised learning approach used for both regression and classification tasks, which seeks to find an optimal separation hyperplane that allows input vectors to be separated in classes [Cortes and Vapnik 1995].

Since the input vector is nonlinearly separable, SVM maps the input vector into a high dimensional feature space (where the separation hyperplane is constructed) through nonlinear mapping. This mapping is done by a kernel function, which defines the hyperplanes and depends only on the input vector space.

In this work, three kernel functions are evaluated in the SVM classifier: linear, polynomial (degrees 1, 2, and 3), radial base function (RBF), and sigmoid. The SVM approach examines the penalty (or regularization) parameter C in the range $1 - 10$, in increments of 1, and in the values 100 and 1000, which aims to penalize the error during the error minimization process made during SVM training. For approaches RBF, polynomial and sigmoid, the kernel coefficient (gamma) is investigated, whose value is $1.0e^{x}$, x ranging from -9 to $+3$, a total of 13 values, the fourteenth value is obtained through $\frac{1}{n\_features}$, where $n\_features$ is the number of features, as shown 3.4.2.

3.5 Data augmentation

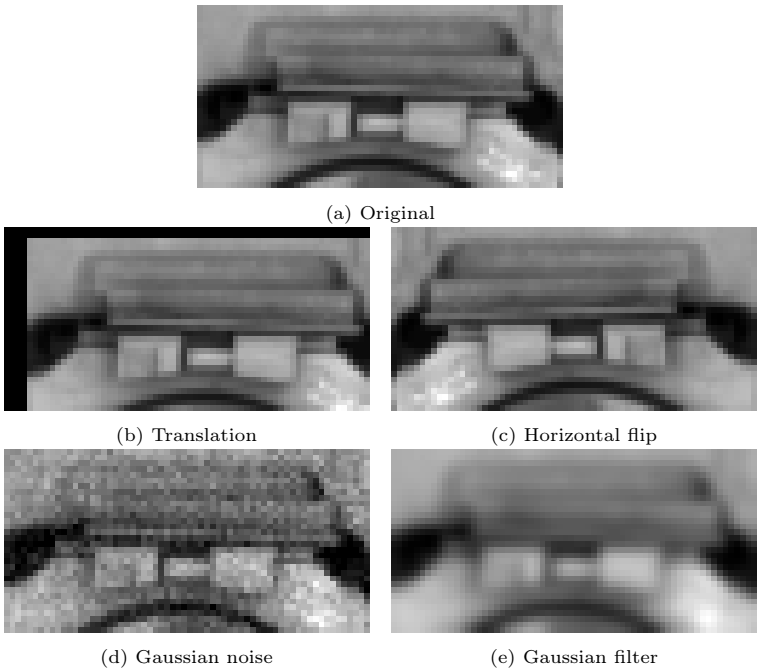
To assess whether the number of training images of the model is sufficient to achieve a good classification performance, data augmentation is used, which artificially expands the training set samples number through transformations, disturbances, or noises in the images preserving their origin class [Chatfield et al. 2014]. Furthermore, data augmentation is commonly used to reduce the overfitting presented during neural network training [Krizhevsky et al. 2012].

Translations are done randomly between -3 and 3 pixels on the horizontal and vertical axes. Gaussian noise, on the other hand, has a fixed mean of 0 and standard deviation varying between 0.001 and 0.05, values that are used in the normal (Gaussian) distribution that generates noise applied to images. Finally, the Gaussian kernel used in convolution with the image that will result in the blurred image is generated by a standard deviation ranging from 0.3 to 1.

Fig. 5 shows the transformations through the data augmentation used in the images of the pad. Fig. 5a shows the original image present in the dataset, which reflects the image of a damaged pad (class 3). Figs. 5b– 5e show the transformations carried out on the original image (Fig. 5a), which are: translation of 3 pixels on the positive horizontal axis and 2 pixels on the positive vertical axis is shown in Fig. 5b; horizontal flip (Fig. 5c); Gaussian noise of zero mean and standard deviation of 0.05 to generate disturbances to the image is shown in Fig. 5d; and Fig. 5e shows the use of a Gaussian filter with a value of 1 standard deviation of the Gaussian kernel to blur the image.

3.6 Configurations

The proposed approach (CNN) has 50 training epochs, with a learning rate of 0.001 and batch size of 16. While the training conducted by comparative methods (SVM and MLP) is 1000 epochs.



(a) Original

(b) Translation

(c) Horizontal flip

(d) Gaussian noise

(e) Gaussian filter

Fig. 5: Transformations are performed by data augmentation. The component image (a) is transformed by translation (b), horizontal flip (c), Gaussian noise (d), and Gaussian filter (e).

The first experiments deal separately and the combination of component images of the space (α) and frequency (β and γ) domains, where the method $\alpha\beta$ represents which original image (spatial) and the DFT magnitude (frequency) are used as input. These experiments use hold-out validation with 80% of the images dataset for training and 20% for the test of the generated model, where the results represent the average of 100 trainings for each of CNN's input combinations.

The parameters that best fit the MLP and SVM models are investigated with K-fold cross-validation, so those with the best performance for each optimizer and kernel are presented. In the work is used $K = 5$, where the results presented represent the mean of 5-folds.

The best fit settings for MLP and SVM are then compared to the proposed approach through hold-out validation. Finally, is discussed an investigation of data augmentation to all approaches.

In data normalization, the mean μ and the standard deviation σ from Eq. 1 are obtained from the training set and used to normalize the feature vector in the training and test sets.

The evaluation of the experiments is performed by the accuracy (A) that describes the overall effectiveness of the classifier, as shown in Eq. 2, where TP and TN are the correctly classified examples of the positive (P) and negative (N) classes, and FP and FN are the wrongly assigned examples of the P and N classes [Sokolova and Lapalme 2009].

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

The concept of positive and negative classes are commonly used in the binary classification context, but in the work, the multiclass classification is analyzed, thus, these concepts are slightly different. The positive class represents the particular class currently analyzed and the negative class represents all other classes.

3.7 Statistical tests

To statistically compare the approaches, specifically the results obtained through the hold-out validation of 100 executions, statistical tests are conducted to analyze the hypothesis that the results are statistically equal. For this, it is necessary to follow a test sequence considering the 5% significance level, as shown below:

(1) It is performed the Grubbs's test to investigate the hypothesis of outlier samples in a data distribution. If the hypothesis of the sample to be outliers is not rejected, this sample is then removed for the subsequent normality test.
(2) To conduct the normality test, the Shapiro-Wilk test is carried out, which investigates the hypothesis that a data distribution comes from a Gaussian distribution (normal). If it is not possible to reject the hypothesis, it is possible to perform parametric statistical hypothesis tests to compare the approaches, such as the analysis of variance test.
(3) To perform the analysis of variance test, it is necessary that the distributions are Gaussian and have the same variance. Thus, it is necessary to perform the Levene's test, which investigates the hypothesis that different distributions have equal variance.
(4) The analysis of variance test (ANOVA) investigates the hypothesis that the means of different distributions (two or more) are the same.
(5) Finally, is performed the Tukey test between pairs of distributions, to assess the significant difference between the means of two distributions.

4. RESULTS AND DISCUSSIONS

Table I presents the results of the images of the spatial (method α) and frequency (methods β and γ) domains of the pad, in addition to the combination of DFT magnitude and phase, appointed by $\beta\gamma$.

Table I: Performance of the classification performed using only the magnitude and/or phase of the DFT of the component image as a network input compared to the use of the original images. The evaluation metrics are accuracy from the test set.

Method	α	β	γ	$\beta\gamma$
A (%)	93.43	84.81	79.15	87.89

The classification based on the original images in the gray-scale was the one that obtained the best result, reaching an accuracy of 93.43%, as shown in Table 1. Then, the combination of DFT products (magnitude β and phase γ) reached an accuracy of approximately 87%, indicating that the frequency information can be a source of sufficient patterns to identify the desired objects.

In isolation, the phase (method γ) information was the one with the worst performance among the methods (Table I), with an average error of approximately 83 of the 396 test images. The phase information of DFT in images is a very complex source of information for the identification of patterns in isolation.

The magnitude identifies, for example, the presence or absence of edges in the image. From Figs. 2d-2f it can be seen that the pad images have components of all frequencies at different levels, and that, in general, the magnitude becomes smaller at high frequencies. The images also show that there are dominant directions in the image that runs through the center of the transformed image, representing the existence of regular patterns in the original image.

The value of each point in Figs. 2g-2i determines the phase of the corresponding frequency. Although it is possible to identify the vertical and horizontal lines corresponding to the patterns in the original

image, the phase image does not produce much new information about the image structure of the spatial domain.

Fig. 6 illustrates the distribution of test accuracy given methods β and $\beta\gamma$, those that resulted in accuracy greater than 80% from DFT information. Concerning the method $\beta\gamma$, the results indicated a lesser dispersion between their executions, as well as the occurrence of a single outlier that exceeds the upper limit (0.9191) of the boxplot. The method β showed greater variability, in addition to two outliers, one of them with the lowest accuracy of 0.5883.

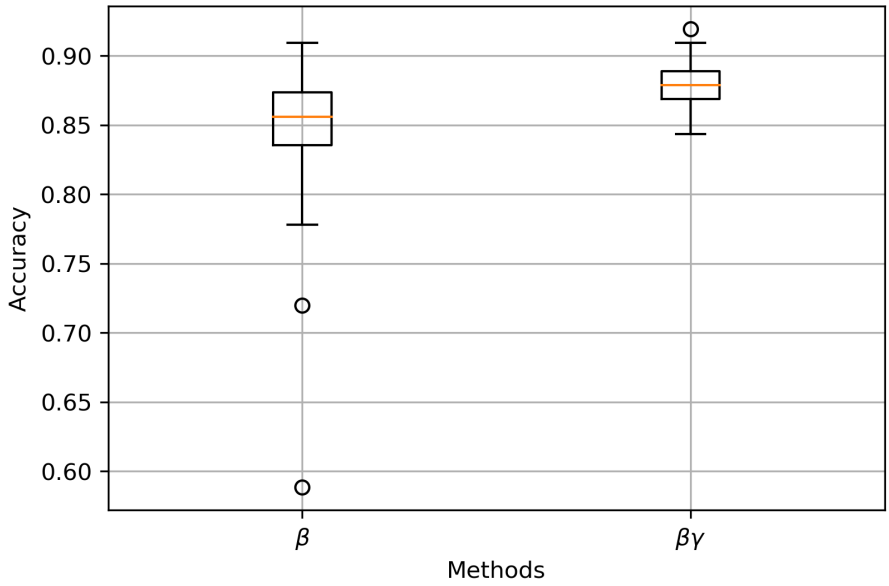


Fig. 6: Comparative boxplots with the result of 100 experiments carried out with methods β and $\beta\gamma$.

Table II presents the results with images from both domains, spatial and frequency, in comparison, again, with the original image. In general, the union of the frequency information with the original image surpassed the accuracy of the tests with the magnitude or phase, alone or combined.

Table II: Performance of the classification of images combined between spatial and frequency information. For comparative purposes, the test results with the original images, method α, have been added to the table.

Method	α	$\alpha\beta$	$\alpha\gamma$	$\alpha\beta\gamma$
A (%)	93.43	94.41	93.32	94.36

The methods $\alpha\beta$ and $\alpha\beta\gamma$ showed a mean accuracy above 94%, and the combination of the original image and its magnitude (method $\alpha\beta$) reached the maximum global accuracy (94.41%).

The boxplots in Fig. 7 illustrate the results of experiments with methods α, $\alpha\beta$ and $\alpha\beta\gamma$. It is possible to observe the similarity between the responses obtained during the execution of the methods $\alpha\beta$ and, $\alpha\beta\gamma$ that do not present outliers. Besides, the median value (orange line) is quite similar between them.

Of the possible accuracy values, the use of information from the original images and their magnitudes (method $\alpha\beta$) vary from just under 92% to just over 97%, this shows that there are still adjustments that can be made to improve this performance.

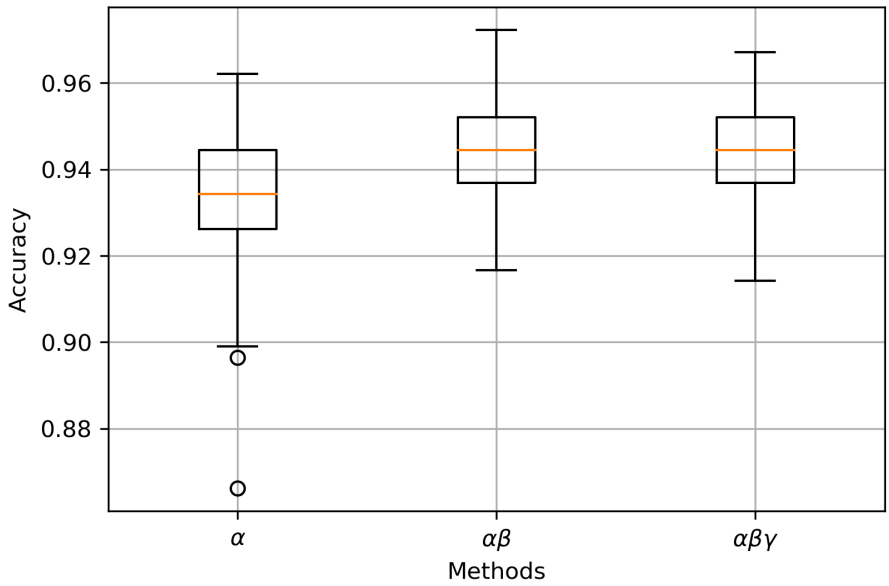


Fig. 7: Comparative boxplots with the result of 100 experiments performed with the best performances in Table II. The three boxplots illustrate the tests with the original images (α), the combination of the original images and the magnitude ($\alpha\beta$) and, the combination of the original images, their magnitudes, and phases ($\alpha\beta\gamma$).

It is performed a statistical analysis of the results in Table II as described in section 3.7. The Shapiro-Wilk normality test informs us it is not possible to reject the hypothesis that the accuracy distributions come from a Gaussian distribution. Performing the Levene's test it is indicated that the hypothesis of equality of variances between the methods cannot be rejected.

ANOVA indicates that it is possible to reject the hypothesis of equality between the means of the accuracy distributions in Table II. Finally, is performed the Tukey test to compare the significant difference between the means of two distributions. The Tukey test to comparisons $\alpha\gamma-\alpha$ and $\alpha\beta\gamma-\alpha\beta$ shows us that we cannot reject the hypothesis of equality accuracy.

The investigation of the comparative approaches (MLP and SVM) are carried out by feature extraction from the HOG descriptor of the original image α (spatial domain) and the image of the DFT magnitude β of the analyzed component, the pad.

It is performed a search to find the parameters that best fit the MLP and SVM models, the investigated parameters were presented in 3.4.3 and 3.4.4. Tables III and IV present the parameters that obtained the best performance of the optimizers and kernels of the MLP and SVM approaches, respectively.

Table III: Summarizing the results of the optimizers and activation functions that best fit the MLP method. The results represent the mean and standard deviation of the 5-fold cross-validation.

Optimizer	Activation function	A (%)
L-BFGS	ReLU	91.49 (± 1.68)
SGD	Logistic	90.89 (± 1.81)
Adam	ReLU	91.34 (± 1.80)

In Table III, the L-BFGS optimizer is the one that achieved the best performance among MLP approaches, where the ReLU activation function was the one that best adjusted to this optimizer, obtaining an accuracy of 91.49%, in addition to the lowest standard deviation (1.68%) between k-fold results. Then, it is possible to notice the behavior similar to L-BFGS presented by Adam (Table III),

having the same activation function as the most accurate method. The activation function best fits the SGD optimizer is logistics, among the three activation functions investigated, however, this is the least performing method, reaching only a mean of 90.89% and a standard deviation of 1.81% accuracy.

Investigating the search for the best SVM parameters in Table IV, it is noted that the worst performance in terms of accuracy is presented by the linear kernel with penalty term $C = 1$, with 90.43%. The polynomial ($C = 3$ and degree 3) and sigmoid ($C = 2$) kernels present intermediate performance between the results, with accuracy lower than 92%. Among the kernels and set of values ($1 - 10$, 100 and 1000) of the penalty parameter, the one that achieved the best performance of the SVM was RBF with $C = 2$, reaching a mean accuracy of 93.01%. Most results from Table IV are obtained models with a kernel coefficient value $\frac{1}{n\_features} = 6.61e^{-4}$ (as shown in 3.4.4), which best fits the three kernels (polynomial, RBF and, sigmoid).

Table IV: Summarizing the results of the kernels and penalty parameter values that best fit the SVM method. The results represent the mean and standard deviation of the 5-fold cross-validation.

Kernel	C	A (%)
Linear	1	90.43 (± 2.30)
Polynomial	3	91.49 (± 1.75)
RBF	2	93.01 (± 1.33)
Sigmoid	2	91.09 (± 1.23)

Both the MLP of L-BFGS optimizer and the ReLU activation function and the SVM of RBF kernel and $C = 2$ are compared to the CNN of α and β inputs, the proposed approach. However, to evaluate the approaches fairly, MLP and SVM are performed by hold-out validation with 100 repetitions, just like CNN. Table V shows the comparison between the L-BFGS, RBF and, CNN methods.

Table V shows that the highest accuracy is presented by the proposed approach, CNN, reaching 94.41%, while the worst accuracy is presented by the LP-BFGS optimizer MLP, with only 92.41% of classification accuracy, that is, 365 correct images out of a total of 396.

Table V: Comparison between the results of the proposed CNN approach and the comparative methods L-BFGS, and RBF.

Method	L-BFGS	RBF	CNN
A(%)	92.41 (± 0.98)	93.75 (± 1.05)	94.41 (± 1.08)

In Table VI the data augmentation (named D) is added to the analysis of the approaches present in Table V. As in Table V, hold-out validation is also used to obtain the results shown in Table VI. Among the three approaches presented, the one with the worst performance, as in Table V, is the D-L-BFGS, with only 93.82%. Meanwhile, the best accuracy performance is presented by the proposed method, which reaches 96.04% (380 images of 396 in the test set) and has the lowest standard deviation (0.88%) among the results, besides, D-CNN is the one with the most significant increase in accuracy when using data augmentation, with 1.72%.

Table VI: Comparison between the results of the proposed CNN approach and the comparative methods L-BFGS and RBF with the addition of data augmentation in all approaches.

Method	D-L-BFGS	D-RBF	D-CNN
A(%)	93.82 (± 1.03)	94.82 (± 1.03)	96.04 (± 0.88)

After a statistical analysis of Tables V and VI, it is feasible to infer that it is not possible to reject the hypothesis of the accuracy distributions coming from the Gaussian distribution. When analyzing the variance by the Levene's test, both the sets of accuracy distributions in Tables V and VI concludes,

it is not possible to reject the hypothesis that the set of distributions in each of the Tables V and V, separately, have equal variances.

Concerning ANOVA, the test informs us that it is possible to reject the hypothesis of equality of mean between distributions of the accuracy of each group, separately. When comparing pairs of accuracy distributions in Tables V and VI (separately), the Tukey test indicates that it is possible to reject the hypothesis of equality between the means of the methods for all possible comparisons. In this way, it is possible exemplify statistically the significant difference between the methods (both Tables V and VI), corroborating the indication of the best performance method presented by CNN and D-CNN in relation comparative methods, as shown in the results of Tables V and V, respectively.

5. CONCLUSION AND FUTURE WORK

This work aims to use spatial domain and frequency (DFT) images for automatic inspection of the railcar component, the pad. The proposed approach employs a CNN that uses component images from both domains as input. The comparison is made by the ANN and SVM methods, where the HOG features extraction from the two domain images is performed. Besides, data augmentation is used to investigate classification performance improvement.

The initial results demonstrate a significant performance of the insertion of the image of the frequency domain, by the discrete Fourier transform, specifically the combination of the original image and DFT magnitude. However, the phase information alone does not present a significant performance, compared to spatial and magnitude images alone or combined.

Concerning data augmentation, it is possible to notice a significant increase in the mean accuracy in all approaches, concluding that this is an adequate technique to improve the classification performance. The proposed approach demonstrated a significant increase in classification improvement using the data augmentation when comparing it to the standard CNN approach. Besides, CNN's feature extraction proved to be superior to HOG's, according to the accuracy results than SVM and ANN approaches. In which statistical tests prove the significant difference between the approaches, confirming the best performance presented by the proposed approach.

In future works, we intend to investigate, for comparison, texture extraction techniques such as discrete wavelet Transform, gray levels co-occurrence matrix, and local binary patterns, as well as classifiers such as k-nearest neighbors and logistic regression. Also, we intend to address the use of dimensionality reduction techniques such as principal component analysis and linear discriminant analysis.

Acknowledgment

The authors thank the financial support of the SENAI Institute of Innovation on Minerals Technology (ISI-TM), as well as the logistical support of Vale Institute of Technology (ITV) and Vale S.A. to obtain the dataset used by this work.

REFERENCES

Avriel, M. *Nonlinear programming: analysis and methods.* Courier Corporation, Mineola, New York, 2003.

Bishop, C. M. *Pattern Recognition and Machine Learning.* Information science and statistics. Springer, New York, USA, 2006.

Cha, Y.-J., Choi, W., Suh, G., Mahmoudkhani, S., and Büyüköztürk, O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering* 33 (9): 731–747, 2018.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

Cortes, C. and Vapnik, V. Support-Vector Networks. *Machine Learning* 20 (3): 273–297, 1995.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, San Diego, CA, USA, 2005.

Gibert, X., Patel, V. M., and Chellappa, R. Deep multitask learning for railway track inspection. *IEEE Transactions on Intelligent Transportation Systems* 18 (1): 153–164, 2017.

Haidari, A. and Tehrani, P. H. Thermal load effects on fatigue life of a cracked railway wheel. *Latin American Journal of Solids and Structures* 12 (6): 1144–1157, 2015.

Han, J., Kamber, M., and Pei, J. *Data mining: concepts and techniques*. Morgan Kaufmann - Elsevier, 225 Wyman Street,Waltham, MA 02451, USA, 2011.

Hart, J. M., Resendiz, E., Freid, B., Sawadisavi, S., Barkan, C. P. L., and Ahuja, N. Machine vision using multi-spectral imaging for undercarriage inspection of railroad equipment. In *Proceedings of the 8th World Congress on Railway Research, Seoul, Korea*. WCRR, Seoul, Korea, 2008.

IWnicki, S. CRC Press, Boca Raton, pp. 548, 2006.

Juang, C.-F. and Chang, C.-M. Human body posture classification by a neural fuzzy network and home care system application. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37 (6): 984–994, 2007.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, Conference Track Proceedings*. ICLR 2015, San Diego, CA, USA, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. Curran Associates Inc, New York, USA, pp. 1097–1105, 2012.

Macucci, M., Di Pascoli, S., Marconcini, P., and Tellini, B. Derailment detection and data collection in freight trains, based on a wireless sensor network. *IEEE Transactions on Instrumentation and Measurement* 65 (9): 1977–1987, 2016.

Odanovic, Z. Analysis of the railway freight car axle fracture. *Procedia Structural Integrity* vol. 4, pp. 56–63, 2017.

Park, B., Chen, Y. R., Nguyen, M., and Hwang, H. Characterizing multispectral images of tumorous, bruised, skin-torn, and wholesome poultry carcasses. *Transactions of the ASAE* 39 (5): 1933–1941, 1996.

Sakhare, K., Kulkarni, A., Kumbhakarn, M., and Kare, N. Spectral and spatial domain approach for fabric defect detection and classification. In *2015 international conference on industrial instrumentation and control (ICIC)*. IEEE, United States, pp. 640–644, 2015.

Samant, N. and Sonar, P. Mammogram Classification in Transform Domain. In *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, Noida, Delhi-NCR, pp. 56–62, 2018.

Sokolova, M. and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* vol. 45, pp. 427–437, 2009.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering* 63 (7): 1455–1462, 2016.

Tao, Y., Muthukkumarasamy, V., Verma, B., and Blumenstein, M. A texture extraction technique using 2D-DFT and Hamming distance. In *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003*. IEEE, Xi'an, China, pp. 120–125, 2003.