IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 6, 2021, accepted November 30, 2021, date of publication December 22, 2021, date of current version January 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3137633

A New Interpretable Unsupervised Anomaly Detection Method Based on Residual Explanation

DAVID F. N. OLIVEIRA[1], LUCIO F. VISMARI[1], ALEXANDRE M. NASCIMENTO[1,2], JORGE R. DE ALMEIDA, JR.[1], PAULO S. CUGNASCA[1], JOÃO B. CAMARGO, JR.[1], LEANDRO ALMEIDA[3], RAFAEL GRIPP[3], AND MARCELO NEVES[3]

[1]School of Engineering, University of São Paulo (USP), São Paulo 05508-900, Brazil
[2]Stanford University, Palo Alto, CA 94305, USA
[3]VALE S.A., Vitoria, Espírito Santo 35400-000, Brazil

Corresponding author: David F. N. Oliveira (davidoliveira@alumni.usp.br)

This work was supported by VALE S.A. under Grant 4600043577.

ABSTRACT Despite the superior performance in modeling complex patterns to address challenging problems, the black-box nature of Deep Learning (DL) methods impose limitations to their application in real-world critical domains. The lack of a smooth manner for enabling human reasoning about the black-box decisions hinder any preventive action to unexpected events, in which may lead to catastrophic consequences. To tackle the unclearness from black-box models, interpretability became a fundamental requirement in DL-based systems, leveraging trust and knowledge by providing ways to understand the model's behavior. Although a current hot topic, further advances are still needed to overcome the existing limitations of the current interpretability methods in unsupervised DL-based models for Anomaly Detection (AD). Autoencoders (AE) are the core of unsupervised DL-based for AD applications, achieving best-in-class performance. However, due to their hybrid aspect to obtain the results (by requiring additional calculations out of network), only agnostic interpretable methods can be applied to AE-based AD. These agnostic methods are computationally expensive to process a large number of parameters. In this paper, we present the RXP (Residual eXPlainer), a new interpretability method to deal with the limitations for AE-based AD in large-scale systems. It stands out for its implementation simplicity, low computational cost and deterministic behavior, in which explanations are obtained through the deviation analysis of reconstructed input features. In an experiment using data from a real heavy-haul railway line, the proposed method achieved superior performance compared to SHAP, demonstrating its potential to support decision making in large scale critical systems.

INDEX TERMS Autoencoder, explainability, fault diagnosis, interpretability, safety.

I. INTRODUCTION

In recent years, advances in deep learning (DL) have enabled themselves to be widely adopted in real-world applications. As a prime representative, the Deep Neural Networks (DNN) have been achieving impressive performance and reaching the state-of-the-art performance in diverse application domains such as computer vision, speech recognition and natural language processing [1]. These achievements drove Machine Learning (ML), in particular the DL, to become a growing research topic due to its potential in automating activities dependent on higher levels of cognition previously performed only by humans.

The application of DL techniques leads to structurally complex models. These complex structures preclude intuitions about the rationale behind the predictions, undermining the model trust. Consequently, DL-based models are adopted as black-box approaches by their end-users. However, this lack of transparency on how the results are obtained stand as a challenge for the adoption of these models in some real-world applications. Depending on the domain, some aspects are critical, such as safety (trust by robust reasoning in risky situations); fairness (equal opportunities ensuring no gender bias or racism) [2]; and accountability (capable to justify the decision making). Besides, critical systems demand their

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaojun Steven Li.

decision-making models to learn and be able to apply reasonable rules in order to avoid unacceptable behavior that may lead to ethical/legal issues, profit losses or accidents.

Therefore, understanding how a decision is made by a ML-based model is fundamental in critical domains. For these reasons, Interpretable Machine Learning has received growing attention by the scientific community, given that it deals with the intrinsic transparency limitations in black-box models. Interpretability can be defined as the ability to enable intuition and reasonability about model output [3] as a means to augment trust [4]. Consequently, interpretable models are a class of methods that use ML-based models to extract relevant knowledge from the data [5], enabling clearness and insights about results obtained from black-box models.

In general, interpretability methods are based on either the perturbation of the model's inputs (perturbation-based) or the analysis of the model's neuron activation/gradient behavior (gradient-based, exclusive for neural networks). Perturbed-based methods evaluate the impact on the output regarding different combinations of deviated samples in a particular input. The common manner to obtain explanations is by building *proxy models* [6], that is, a shallow model (such as linear regression or decision tree) that locally approximate the black-box behavior in the region of the particular input to be predicted. This shallow model is easily interpretable due its simple structure. Thus, perturbation-based methods are agnostic with respect to the predictive models by only requiring access to the model's boundaries (input-output tuple), enabling them to be applied to any ML-based model.

However, the trade-off between interpretation performance and response time is a inherent challenge due to the combinatorial aspects in order to find optimal results. In cases where the model input has a large number of parameters, one single record may require a large processing time to achieve more accurate interpretation results. In the same way, faster results can be obtained by sacrificing interpretation performance due to the randomness brought by searching for results in a smaller dimensional space. Consequently, perturbed-based methods are prohibitive in many real-time, large-scale applications. Examples of well-known perturbed-based methods are LIME (Local Interpretable Model-agnostic Explanations) [7] and SHAP (SHapley Additive explanation) [8].

On the other hand, the gradient-based methods analyze neural network behaviors by looking for which network nodes are more excited (or deviated from the normal behavior), given a particular output. Gradient-based approaches are not negatively impacted by the size of inputs since they only evaluate a specific observation, resulting in a fast response. Furthermore, these methods deliver the exact same interpretation result for a given network and to the same input/output (deterministic behavior). Consequently, they are more suitable for real-time, large-scale purposes, allowing them to obtain robust interpretation results during runtime. Examples of this technique are DeepLIFT (Deep Learning Important FeaTures) [9] and Integrated Gradients [10].

Despite gradient-based methods being more suitable for real-time large-scale models, they are not agnostic, staying restricted to neural networks (i.e. DL-based models): The input-output relationship is raised by checking detailed behaviors of neuronal activation during predictions. In all other ML-based applications, the perturbed-based methods are the only option for interpretability, which is prone to computationally inefficiency to deal with a large number of parameters.

A very typical domain in which ML is extensively applied in large scales systems is Anomaly Detection (AD). AD is used to identify items or events that deviate from an expected behavior [11]. AD is a common challenge in diverse domains, such as finance (fraud detection), network security (intrusion detection), industrial (fault detection) and healthcare (patient health checking). Independent of the application domain, ML-based approaches need to be trained to learn what is an anomaly, in which can be found diverse approaches using supervised, unsupervised or semi-supervised adapted for to this task.

Regardless the superior performance of supervised methods training ML-based models, AD imposes restrictive training requirements concerning to the task specificity's, such as (i.) supervised approaches are prone to be biased to mask anomalies, due to anomalies are rare events [12] and, consequently, the datasets are very imbalanced; and (ii.) anomalies may happen in a novel and unseen ways [13], meaning that the assumption that all possible anomalies are known and mapped is rarely true. Therefore, supervised approaches may be inadequate and fail prone when there is limited knowledge (i.e. insufficient number of examples or lack of all possible anomaly scenarios for training). On the other hand, by exploring deviations in data, unsupervised AD approaches are flexible and more efficient to detect anomalies, including the cases when the observations come in a novel form. Additionally, depending on the domain problem the data might be collected in different formats, wherein there are adequate AD-based models to deal with contextual anomalies in images (spatial anomaly) [14] or time-series (temporal anomaly) data [15].

Autoencoder (AE), a type of neural network architecture, is the core of current unsupervised DL-based models for AD applications [13], in which a substantial number of researches can be found for AD issues. Unsupervised AD applications using AEs are composed by two basic components: (i.) The AE (neural network), generating a deviation score (a.k.a. residuals) to the data input; (ii.) A binary classifier, which decides about the input value (true or false) based on AE residuals rates. Thus, given the AD's Input-Output relationship is not completely implemented by a NN, the results obtained by unsupervised AD applications using AEs cannot be explained (interpreted) by gradient-based methods. Consequently, the only option is using agnostic interpretable models, such as LIME or SHAP, which faces randomness and time processing challenges.

Therefore, this paper proposes a method to implement interpretability capabilities to AE-based unsupervised

Anomaly Detection (AD) applications. This method stands out for its implementation simplicity, low computational cost, and deterministic behavior. Its explanations are obtained by means of the deviation analysis of reconstructed input features. Consequently, it is more suitable for AE-based AD applications in large-scale systems. We focus our study on tabular data (no spatial/temporal anomaly), which still represent a vast portion of the data available for anomaly detection applications in real-world problems.

This paper is organized as follows: section 2 gives a background information for the present research; section 3 presents our proposed method; section 4 presents an experiment to validate the proposed method in a real-world critical application; section 5 presents the concluding remarks of this work, as well as its possible future steps.

II. BACKGROUND

This section provides background knowledge about keys concepts supporting this work: section 2.1 gives a brief overview about reconstruction-based methods and traditional design of AE for AD tasks; section 2.2 describe a intuition about SHAP (kernel explainer - the baseline method compared with RXP) and LIME (interpretability method extended by SHAP); section 2.3 highlights some related works that propose different approaches to deliver interpretations about their results.

A. AUTOENCODER FOR UNSUPERVISED AD (AE-BASED AD)

Different manners are found in the literature to deal with AD problems, such as distance-based [11], density-based [16], or custom heuristics [12]. Among them, reconstruction-based approaches have gain attention in recent researches using DL architectures [13].

Reconstruction-based approach attempts to transform the data into different dimensions (usually smaller), and then inversely returning to the original format. As anomaly detection problems usually deal with unbalanced datasets (since anomalies are "rare events") [12], the philosophy of using reconstruction-based models understands that models are trained to reproduce normal observations with high performance. On the other hand, these models tend to fail to faithfully reproduce anomalies. Thus, these models discriminate samples by calculating a residual error between normal (usually lower values) and failure (higher values). Common examples of back-end methods for reconstruction-based models are Autoencoders (AE), Restricted Boltzmann Machines (RBM), and Principal Component Analysis (PCA).

As a prominent example of DL architecture applied for AD problems, Autoencoder (AE) is a type of artificial Neural Network (NN) that learns to reconstruct the network input to its output using unsupervised training. This is typically done by compressing the input through lower dimension layers (encoder), and then decompressing (reconstructing) the original input to its original size (decoder). This architecture allows AEs to be used for efficient Dimensionality Reduction

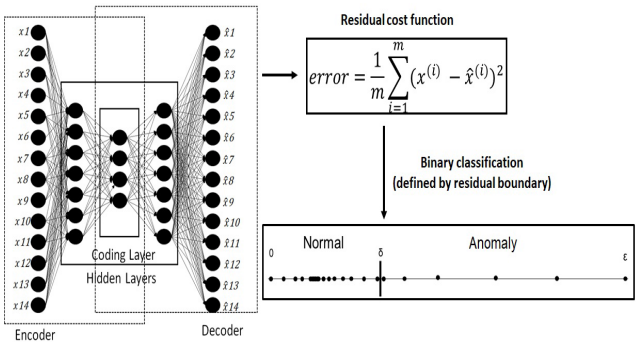


FIGURE 1. Traditional approach for standard autoencoder applied to anomaly detection. Source: authors.

(DR) and generative models. AEs learn intrinsic aspects from the data, creating a reduced dimensional latent space that discards noises and unrepresentative data.

Figure 1 illustrates a traditional AE implemented as a feed-forward Multi-Layer Perceptron, typically applied to anomaly detection (AD). A multidimensional input is processed by artificial neurons through a sequence of nonlinear functions over multiple layers. This combined with the dimensional reduction over the layers results into rich dense representations in its hidden layers. The compressed data is then used to rebuild the original information from the lower representation. Many variations of AE can be found in the literature, such as Sparse [17], Denoising [18], Variational [19], Adversarial [20] and hybrid combinations such Recurrent [21], [22] and Convolutional [23].

Since AE for AD follows the reconstruction-based approach, are classified by calculating residual error in the reconstruction of the encoded input by the decoder. Residual values are commonly obtained by a mean squared error function, wherein for each parameter i, x_i is the original input, $\hat{x}_i$ is the reconstructed input. A contamination hyperparameter (percentage of faulty records in the data) should be defined to find the residual boundary value between normal and fault for the binary decision. During the training phase, the lowest record within the highest residual contamination percentage defines the residual threshold value (δ) for the binary classification between normal and anomaly.

B. AGNOSTIC INTERPRETABILITY METHODS

Among the manners to provide interpretations from black-boxes, model-agnostic interpretability methods stand out for the independence about intrinsic aspects of the predictive model to produce responses. Interpretations are commonly generated by forcing noises in the input and checking respective impacts on the outputs. Higher deviations may indicate that the perturbed attribute is more relevant to the particular prediction. Below we detail LIME e SHAP (kernel explainer), two of the most popular methods in this line.

1) LIME

LIME [7] grounds on the assumption that complex black-box models can be simulated by shallow models in the region of

the instance to be explained (also known as local surrogate methods). A white-box model is trained using generated perturbed samples around the original input. The optimization of the white-box method aims at maximum approximation as to the response of the black-box model in the region of the observation to be explained, which can be expressed as follows:

$$e(x) = \underset{g \in G}{argmin}\, L(f, g, \pi_x) + \Omega(x) \tag{1}$$

wherein the function e represents the best explanation for the reduction of the residual error L (usually the quadratic error), between the black-box model f and the approximate white-box model g, belonging to the space of possible explanations G, with control of parameters by regularization Ω, and trained with a sample volume of neighbors π_x (the more neighbors, the greater the quality of results). The white box methods are generated with the objective of locally approximating the prediction results of the black box methods, and provide intuitive means of interpretation, native for their structure, such as linear regression, which follows the form

$$g(x) = \beta_0 + \sum_{i=1}^{M} \beta_i x_i \tag{2}$$

in which the input x composed of M parameters and their respective weights β (in which β_0 is the bias) are added. The parameter importance is assessed by of the module its value associated with weight ($\beta_i x_i$). Other types of white-box method can be applied, such as decision trees, and decision rules (IF-THEN). In addition, the method is prepared to receive different types of entries such as tabular records, free text or images.

2) SHAP (KERNEL EXPLAINER)

SHAP [8] is an interpretability method inspired by Shapley Values [24], a solution concept on cooperative game theory. This game comprises of several participants (input features) that act with a common goal to be achieved (prediction), are rewarded fairly in relation to his contribution to gambling (feature relevance), being represented by the function

$$\varphi_i(v) = \sum_{z \subseteq N\{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} \left(v(z) - v(z_i)\right) \tag{3}$$

in which calculates the average contribution (Shapley Value) of a player (i.e. input feature) i in relation to the total number of players (i.e. full set of features) N. This is done by capturing his marginal contribution (i.e. feature relevance) on the difference obtained between his participation or not in different coalition permutations (subgroups) of players z, calculated by function v. The contribution of an ordered coalition of players is represented by $v(z)$, and $v(z_i)$ represents the same coalition without the contribution of player i. The contribution of each player varies depending on the set of players that compose the coalition. A unique distribution of the average contribution of all game participants is created, through checking each of the possible permutations. The Shapley value obtained corresponds to the marginal contribution of each parameter, obtained fairly in relation to the importance of all parameters.

Lundberg ([8]) raises three important properties delivered by SHAP that are consistent concerning Shapley Values, suggesting that the features relevance weighting are done fairly with a strong theoretical background (being a disregarded matter by LIME): (i.) *local accuracy*, which similar to LIME, this method has the objective of local approximation about the behavior of a black box method regarding a specific observation, generating an explanation by the linear additive combination, similar to (2), in form

$$g(x) = \varphi_0 + \sum_{i=1}^{M} \varphi_i x_i \tag{4}$$

in which the Shapley Value (φ_i) represents the weight of parameter i on the prediction of a point observation. (ii.) *lack of importance*, meaning in a simplified view of the reduced dimensional input, the lack of importance indicates the absence of a parameter without affecting the expected response; (iii.) *consistency*, denoting that if the white box model is updated so that the contribution of one parameter increases or remains the same, regardless of the other parameters, its value must not be reduced.

In cases where the input data have an extensive number of parameters leads to a high computational cost over coalition permutations. SHAP suggests an optimized way to calculate important parameters using the conditional average of the original Shapley Values model. The author suggests that the premise of independence and linearity (4) between the parameters simplifies the computational complexity. This is done by reducing the cases to be evaluated (i.e. the evaluation of coalitions of parameters are not evaluated in different orders, only the inclusion of new parameters). Then, it demonstrates that a white box function can be approximated in a more efficient way.

SHAP deals with the error function stated by LIME (1), but defining specific patterns and adjustments on each of the components to minimize the error, which is called the *SHAP Kernel Explainer*. It suggests well-defined methods for optimization (error minimization standards and adjustments generally suggested by LIME), promoting an improvement and efficiency gain for the generation of samples. This paper, Kernel SHAP is used on the experiments. The authors also suggest solutions aimed at specific types of models, which are not discussed in this work.

C. RELATED WORKS ON INTERPRETABILITY METHODS TO AE-BASED AD

Some relevant works dealing with interpretability methods using AE can be found in the literature. [25] propose a combination with SHAP using the ranking of top residual features obtained by the AE to be interpreted individually against the explainer. SHAP values are recalculated using the set of

interpretations. [26] present a method called ALIME, which combines denoising autoencoders and LIME by optimizing the sampling task through the weighting training instances using the latent space representations. Authors show ALIME improves stability on the results with a faster response time.

[27] detail an interpretable AD method on image streaming for automatic dependent surveillance-broadcast in air traffic control systems based on LSTM-Autoencoder. The proposed method is capable to identify frames and regions of the image with high accuracy by checking regions with high residual errors. [28] defines a method based on Generative Adversarial Networks (GAN) called AnoGAN that is trained with normal data to detect anomalies in images of X-ray CT scans. An anomaly interpretation is done by comparing normally generated images with real images with similar distribution, in which anomalous regions are identified by strong dissimilarity forms.

Reference [29] use Variational Autoencoders (VAE) for network intrusion detection. Interpretations are extracted by applying clustering methods over gradient behavior on the latent representations, in which anomalies present different patterns from the normal. Reference [30] propose the use of AE and Long Short-Term Memory (LSTM) neural networks for fault detection and diagnosis. AE captures incipient anomalous behaviors and its compressed layer outputs are used as dimensionality reduction strategy for LSTM networks on fault diagnosis tasks.

Thus, diverse AE architectures to compose interpretable models can be found in literature. Some works just use AE as auxiliary tools to optimize other predictive methods [30]. Others use agnostic methods such as LIME [26] and SHAP [25]. Some other works focused on specific architectures such as AnoGAN [28] and LSTM-AE [27], made for specific application domains (spatial and temporal data input types). However, we can observe that all these propositions are prone to dimensionality constraints and, consequently, they are computationally expensive when a large number of parameters is used to as input to AD application. This lack of efficiency can jeopardize their practical use in large-scale AD applications.

Therefore, interpretability methods to be used in AE-based AD applications that deal with abundant data availability with low computational cost are demanded. Thus, we propose a new interpretability method that deals with common tabular data (as usual in real industrial scenarios), being capable of delivering deterministic and near real-time interpretations, which makes it more suitable for practical purposes.

III. THE RESIDUAL eXPlainer (RXP) METHOD

Considering the training set $T \in \mathbb{R}^{N \times M}$ having N instances and M features, statistical data (standard deviation and mean) are gathered from the reconstructed residuals for each parameter. This statistical data is used to calculate the Z-score, defined by the Eq.(5), in which z_{nm} is the deviation score for each parameter m belonging to the sample n, x_{nm} is the original input, u_m is the residual training mean, and σ_m is the residual training standard deviation.

$$z_{nm} = \frac{x_{nm} - u_m}{\sigma_m} \tag{5}$$

During the test phase, given one particular instance n classified as anomaly, the relevance R_{nm} of a feature $m \in M$ is defined by Eq.(6), in which the nominator is the residual cost function (squared error between the original x_{nm} and reconstructed x'_{nm} input) weighted by log module of Z-score z_{nm} between the original feature input and its mean residual training value. This weighting module is smoothed by log to reduce the strong effect caused by binary features (when existing together with continuous features) that may lead to very imbalanced relevance attributions. The denominator acts as normalization factor along all feature dimensions. Higher values for R_{nm} means that the attribute m has higher relevance for the decision of the instance n as an anomaly.

$$R_{nm} = \frac{\log(1 + |z_{nm}|)\left(x_{nm} - x'_{nm}\right)^2}{\sum_{i=1}^{M} \log(1 + |z_{ni}|)\left(x_{ni} - x'_{ni}\right)^2} \tag{6}$$

Note this method takes into account not only the deviations on the reconstruction, but considers the deviated input as well (raised by Z-score). Therefore, even for cases in which the network learns good reconstruction of some anomaly, the Z-score part tends to penalize deviated inputs relative to the training data. Consequently, this helps to mitigate possible mistakes that could hinder the interpretability task due to high-quality reconstructed patterns learned by the AD model. In such cases, the observations commonly have slight deviations from the normal behavior, and the AD model returns very low and balanced residual rates, bringing confusion to deliver accurate relevance weighting.

IV. EXPERIMENTAL EVALUATION

AD is a useful approach to detecting faults in industrial applications context, since 'faults' are anomalies – rare events that deviate a system from its expected behavior [31]. Likewise as it has been deeper explored for patient health diagnosis in medical applications [32], we claim that interpretability models can also be useful in the industrial context. This is feasible by delivering further information about the cause of a fault event (fault diagnosis) by raising which parameters were relevant to a fault detection classification. Therefore, our experiments aim to compare RXP against Kernel SHAP as the fault diagnosis solution in a real scenario on machinery operation in the railway domain. In this section we present the case study (dataset), experimental process, evaluation metrics and results.

A. CASE STUDY

We used an exclusive dataset obtained from multiple sensors deployed along the Vitoria-Minas Railway track ('Estrada de Ferro Vitória a Minas' - EFVM, Brazil), operated by the Brazilian mining company VALE S.A.

Three types of wayside equipment were used to get measures from wheels and bearings of rail cars: Hotbox and

TABLE 1. Three examples of parameters from different wayside equipment.

Parameter	Wayside Equipment	Component	Side	Axle
$HEAT-WHEEL-LEFT-AXLE3$	HBW	Wheel	Left	3
$RS-R-AXLE2$	ABD	Bearing	Right	2
$DIR-IMPACT-MAX-AXLE1$	WILD	Axle	-	1

Wheel (HBW) captures thermal emission from wheels and axle boxes (bearings) through infrared sensors; Wheel Impact and Load Detector (WILD) measures the force between the rail and moving wheels through the analysis of vertical load in the line; and Acoustic Bearing detector (ABD) captures acoustic emissions and classify possible faults in wheels and bearings. Totally, 15 HBW, 1 WILD and 1 ABD are deployed along EFVM, and they were used to get data with about 257 parameters.

The raw data, obtained by wayside equipment, contains information at axle level (4 axles per rail car). Table 1 shows an example of parameters captured by each wayside and its respective links to the components and regions on the rail-car. Since many information from the same equipment (rail-car) is spread out over different datasets (waysides) and records (per axles), all records were grouped to rail car level, encompassing all axles in an unique record per track region. This approach enriches the case study by allowing to evaluate the capacity of the method to explain faults in holistic view by checking different causes (overheating, vibration or acoustic), components (wheel, bearing or axle), and regions (left/right side, axle number) of the same rail car.

B. EXPERIMENTAL PROTOCOL

This experiment was performed using a 3-step process, as illustrated in Fig. 2. In **Step (A)**, the AE model were trained for the fault detection task. We followed the hyper-parameter tuning defined by [33]: large number of layers (20); high compression on latent layers (up to 98%); no use of dropout/batch normalization; mean squared error for cost function and residual analysis; hyperbolic tangent as activation for latent layers and sigmoid for output. This model were used to detect faults (binary classification) and their respective residual rates for the diagnosis step (step B).

In **Step (B)**, the interpretability methods is used to obtain the explanations (by means of top-weighted relevant parameters per record) to the faults detected by step A. The proposed method (RXP), Eq.(6), and SHAP [8] were applied and their results are compared in step C. Due to its deterministic behavior (giving the pair input/output), RXP is tested just once per sample set. On the other hand, SHAP may take longer to obtain consistent results when applied in problems with a large number of parameters. Therefore, 3 different scenarios are defined for SHAP, exploring the relationship between results precision and response time. The larger the size of the search space the less random the results, but higher is the response time. On the other hand, the smaller the size of the search space the more prone to unexpected or inconsistent

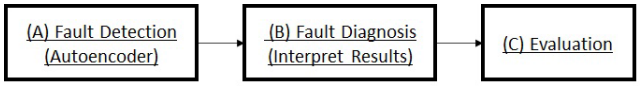


FIGURE 2. The 3-step experimental process.

results, but faster is the response time. Explanation results from this step were evaluated in step C.

Step (C) evaluated the results from each interpretability method. Step B returned a ranking of relevant parameters with their respective weights about the classification. We checked if the expected parameters are ranked in top-weighted, and be capable of evaluating whether the diagnosis returns closer answers relative to the exact expected one. Each interpretability record was evaluated by computing the *Mean Average Precision (MAP) score* (7) over the top relevant parameters. This metric is commonly used in ranking-based problems, and works by recursively raising the precision for each ranking position, starting from the top to the bottom. It captures the concentration of the expected results on the top and penalize any absent expected.

$$MAP = \frac{1}{n}\sum_{n}\sum_{k}(S_{nk} - S_{nk-1})\,P_{nk} \tag{7}$$

S_nk is the recall (sensitivity) and P_nk is precision for the threshold at the top K relevant features for query n. Recall is a metric used to evaluate how capable an AD method is to identify correctly (positively) the 'normal' records in a dataset (i.e. the true positive rate, or TP/(TP+FN)). TP is the number of True Positive results (correctly identified as normal by the AD) - in our case, be part of top-weighted K relevant parameters - and FN is the number of false negative results (relevant parameters wrongly classified as not important). Precision check the frequency of TPs relative to all samples classified as positives (TP/(TP+FP)), in which FP is the false positive cases (not relevant features wrongly classified as relevant).

C. RESULTS AND DISCUSSION

For the fault detection step (A), the AE was trained using around 3.5 million samples in 1 epoch, 32 samples per batch, 25 layers with compression up to 98% between the raw dimension (257) and the coding layer (5). Further details about the other hyperparameters can be found in [33]. The AE was tested against around 379 thousand records containing 3867 faulty samples, achieving a precision (P) of **82.5**% and recall (S) of **95.3**%.

TABLE 2. Experimental results (SHAP1, SHAP2, SHAP3, and RXP).

Overall Results	SHAP1	SHAP2	SHAP3	RXP
MAP (Eq. (7))	80.47%	80.61%	79.54%	81.38%
Mean response time (ms)	11,100	5,650	170	0.272
Paired T-test	3.3×10^{-10}	6.5×10^{-10}	3.2×10^{-12}	-

FIGURE 3. Interpretation results of sample '295' containing 4 fault causes. Source: authors.

The dataset composed by all the faults detected by step A (TP + FP) and false negatives cases (FN) were considered by step B. We performed 30 tests with 200 different random samples (with replacement) in order to reduce any random behaviors obtained from SHAP scenarios. These samples were interpreted by our proposed method (RXP) and SHAP in 3 different scenarios (SHAP1, SHAP2 and SHAP3). SHAP1 was tuned to return more precise results with low responses (around 7 seconds per record) by generating **800** simulated samples using **200** background training examples. SHAP2 was tuned for faster responses (around 5 seconds per record) by generating **800** simulated samples using **100** background training examples. SHAP3 was tuned to deliver very fast results (around 150 milliseconds per record) by generating **80** simulated samples using **10** background training examples, but presenting very unstable interpretation results. Figure 3 shows an example of a true positive sample containing 4 fault causes related to ABD wayside. SHAP1, SHAP2 and RXP were able to identify all the relevant parameters with strong weight importance (comprised by the larger bars on each plot), while SHAP3 fail to identify 2 fault causes.

Results obtained by this experimental process - Mean Average Precision (MAP), Mean Response Time (in miliseconds), and the paired T-test from results (RXP versus SHAP) - are presented in Table 2. Experimental results show that RXP was faster than the SHAP scenarios, achieving up to 625 times faster relative to SHAP3, 20,772 times faster than SHAP2, and 40,845 times SHAP1. Besides, our method achieved the best results regarding the mean average precision (MAP).The RXP delivers superior performance relative to SHAP models, achieving p-value less than 0.00000001% on paired T-test.

V. CONCLUDING REMARKS

In this work, we propose a novel interpretability method for AE-based AD tasks through the reconstruction residual analysis named as RXP. We compared our method with SHAP using real data from railway operations. The experiment compared the methods by simulating a fault diagnosis task, raising the relevant parameters that lead a sample to be classified as faulty (anomaly). For practical purposes, SHAP presented good performance over MAP metric but suffers from the trade-off between delivering consistent results (mitigate random behavior) and cost of evaluating large combinatorial parameter sets. On the other hand, RXP has provided superior results, delivering near-real time response with high precision. Moreover, we focused to check the behavior of RXP over different sample tests relative to SHAP. As a key factor in critical systems, other experiments must be performed by running several tests over the same sample set to highlight possible instabilities caused by the non-deterministic nature of SHAP.

Through the weighted Z-score, our method also becomes sensitive to deviations even on instances with no gross deviation over any particular parameter (but slightly perturbed in a set of them). In this work, we considered the global residual calculation (squared error overall dimensions) and their respective weight of each parameter on the value. Further experiments are demanded in order to check appropriate usage for global residual calculations or relative to the parameter in other scenarios.

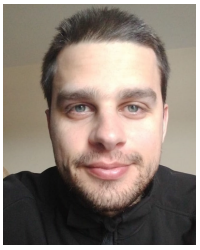
The relevance distribution obtained from the interpretations is a new source of data itself. In this study, we captured and compared the results from only one perspective. We hypothesize that such data may be valuable by allowing new ways to evaluate and complement the analysis. Further studies must be conducted to check possible ways to leverage and discover patterns from it.

Finally, we highlight that our method is independent (agnostic) of the AD approach and may be adapted to any reconstruction-based model, such as Generative Adversarial Networks (GAN) [34], Principal Component Analysis (PCA) [35] and Restricted Boltzmann Machines (RBM) [36] since the method does not require to know the internal architecture of the predictive model. In order to validate the generalization of achievements presented in this work,

future studies may compare RXP with well-known literature benchmarks based on open datasets, and be extended to fit with spatio-temporal anomalies as well.

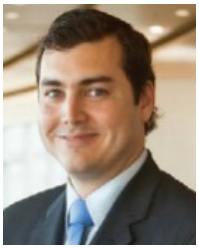
REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.
[3] S. Chakraborty and R. Tomsett, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput.*, May 2017, pp. 1–6.
[4] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
[5] W. James Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: Definitions, methods, and applications," 2019, *arXiv:1901.04592*.
[6] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2018, pp. 80–89.
[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 4765–4774.
[9] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3145–3153.
[10] M. Sundararajan, A. Taly, and Q. Yan, "Gradients of counterfactuals," 2016, *arXiv:1611.02639*.
[11] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, Apr. 2016, Art. no. e0152173.
[12] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.
[13] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.
[14] M. Singh Minhas and J. Zelek, "Anomaly detection in images," 2019, *arXiv:1905.13147*.
[15] A. A. Cook, G. Misirli, and Z. Fan, "Anomaly detection for IoT time-series data: A survey," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6481–6494, Jul. 2020.
[16] T. Hagemann and K. Katsarou, "Reconstruction-based anomaly detection for the cloud: A comparison on the Yahoo! Webscope S5 dataset," in *Proc. 4th Int. Conf. Cloud Big Data Comput.*, New York, NY, USA, Aug. 2020, pp. 68–75.
[17] Z. Chen and W. Li, "Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1693–1702, Jul. 2017.
[18] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.
[19] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[20] S. Rajendran, W. Meert, V. Lenders, and S. Pollin, "Unsupervised wireless spectrum anomaly detection with interpretable features," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 3, pp. 637–647, Sep. 2019.
[21] E. Marchi, F. Vesperini, F. Weninger, F. Eyben, S. Squartini, and B. Schuller, "Non-linear prediction with LSTM recurrent neural networks for acoustic novelty detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–7.
[22] A. Sagheer and M. Kotb, "Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, Art. no. 19038.
[23] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the Neyman–Pearson lemma," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 212–224, Jan. 2019.
[24] L. S. Shapley, "A value for N-person games," in *Contributions to Theory Games*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.
[25] L. Antwarg, R. Mindlin Miller, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using SHAP," 2019, *arXiv:1903.02407*.
[26] S. M. Shankaranarayana and D. Runje, "ALIME: Autoencoder based approach for local interpretability," in *Proc. Int. Conf. Intell. Data Eng. Autom. Learn.* Springer, 2019, pp. 454–463.
[27] S. Akerman, E. Habler, and A. Shabtai, "VizADS-B: Analyzing sequences of ADS-B images using explainable convolutional LSTM encoder-decoder to detect cyber attacks," 2019, *arXiv:1906.07921*.
[28] E. A. Donahue, T.-T. Quach, K. Potter, C. Martinez, M. Smith, and C. D. Turner, "Deep learning for automated defect detection in high-reliability electronic parts," *Proc. SPIE Appl. Mach. Learn.*, vol. 11139, Feb. 2019, Art. no. 1113907.
[29] Q. P. Nguyen, K. W. Lim, D. M. Divakaran, K. H. Low, and M. C. Chan, "GEE: A gradient-based explainable variational autoencoder for network anomaly detection," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Jun. 2019, pp. 91–99.
[30] P. Park, P. D. Marco, H. Shin, and J. Bang, "Fault detection and diagnosis using combined autoencoder and long short-term memory network," *Sensors*, vol. 19, no. 21, p. 4612, Oct. 2019.
[31] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Trans. Depend. Sec. Comput.*, vol. 1, no. 1, pp. 11–33, Jan./Mar. 2004.
[32] R. El Shawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in HealthCare a comparative study of local machine learning interpretability techniques," in *Proc. IEEE 32nd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2019, pp. 275–280.
[33] D. F. N. Oliveira, L. F. Vismari, J. R. de Almeida, P. S. Cugnasca, J. B. Camargo, E. Marreto, D. R. Doimo, L. P. F. de Almeida, R. Gripp, and M. M. Neves, "Evaluating unsupervised anomaly detection models to detect faults in heavy haul railway operations," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 1016–1022.
[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[35] F. R. S. K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philosoph. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.
[36] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.



DAVID F. N. OLIVEIRA received the B.Eng. degree from IESB, in 2009, the post-graduate degree in knowledge management from UFRJ, in 2015, and the M.Sc. degree in computer engineering from the University of São Paulo (Poli-USP), in 2020. Since 2009, he has been acting as a Software Engineer in IT industry in diverse domains. From 2018 to 2020, he was with the Safety Analysis Group (GAS), POLI-USP, as a Researcher in machine learning applied to safety critical systems. Since 2020, he has been working on research and development for recommender engines. His current research interests include interpretability and policy learning applied to recommender systems in cold start problems.

LUCIO F. VISMARI received the B.Eng. and M.Eng. degrees in electrical engineering from the School of Engineering, University of São Paulo (Poli-USP), in 2001 and 2007, respectively. Since 2002, he has been with the Safety Analysis Group (GAS), POLI-USP, as a Senior Researcher in the system safety field. He has more than 40 published scientific works in periodicals, conferences, and books. Besides, he has more than 15 years of experience as an independent safety assessor, working in dozens of large-scale public and private projects in railway, subway, ATM, and defense. His current research interests include the safety assurance challenges in complex engineered systems, mainly the application of cyber-physical systems and new technological trends on high-risk critical domains.



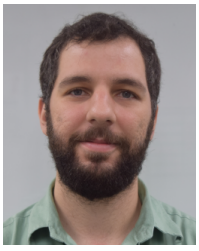
ALEXANDRE M. NASCIMENTO received the B.Eng. degree in mechatronics engineering from the School of Engineering, University of São Paulo (Poli-USP), in 1998, the M.Sc. degree in management with a major in information systems from the School of Management, University of São Paulo (FEA-USP), in 2005, and the M.Sc. degree in management with a major in system dynamics from the Massachusetts Institute of Technology (MIT), in 2013. He is currently a Visiting Researcher at Stanford University. He was at the MIT Media Laboratory for an academic term. He also holds specializations from MIT, the University of California at Berkeley, and Harvard University, in topics such as cognitive robotics, deep learning, image recognition, and text mining. Prior to his current research activities, he worked for Samsung's research and development in Silicon Valley as an International Product Engineering Director for the Samsung Internet of Things (IoT) Platform.

JORGE R. DE ALMEIDA, JR., received the electronic engineering degree from the School of Engineering, University of São Paulo (Poli-USP), Brazil, in 1981, and the M.Sc. and Ph.D. degrees from the Poli-USP, in 1989 and 1995, respectively. He is an Associate Professor with the Computer and Digital Systems Engineering Department, Poli-USP, where he is also a member of the Safety Analysis Group. His research interest includes reliable and safe computational systems for critical applications.

PAULO S. CUGNASCA received the electronic engineering degree from the School of Engineering, University of São Paulo (Poli-USP), Brazil, in 1987, and the M.Sc. and Ph.D. degrees from the Poli-USP, in 1993 and 1999, respectively. He is an Assistant Professor with the Computer Engineering and Digital Systems Department, Poli-USP, where he is also a member of the Safety Analysis Group. His research interest includes area of reliable and safe computational systems for critical applications.

JOÃO B. CAMARGO, JR., received the bachelor's degree in electronic engineering and the M.Sc. and Ph.D. degrees from the School of Engineering, University of São Paulo (Poli-USP), São Paulo, Brazil, in 1981, 1989, and 1996, respectively. He is currently an Associate Professor with the Department of Computer and Digital Systems Engineering (PCS), Poli-USP, where he is also the Coordinator of the Safety Analysis Group (GAS).

LEANDRO ALMEIDA received the B.Eng. and M.Eng. degrees in mechanical engineering from UFES, in 2010 and 2017, respectively. Since 2011, he has been with VALE S.A., working with the development and implementation of the maintenance strategy for passenger car fleets on the Vitória—Minas Railroad (EFVM) and the Ouro Preto-Mariana Tourist Train, as well as the development of materials and suppliers. He is the Technical Responsible for research and development studies and investments aimed at EFVM's rolling stock areas.



RAFAEL GRIPP received the B.Eng. degree in mechanical engineering from UERJ, in 2009, and the post-graduate degree in railroad engineering and the specialization degree in reliability engineering from UTFPR, in 2010 and 2017, respectively. He has been with VALE S.A. as a Railroad Engineer, since 2011. He is working in railroad wagons maintenance with qualitative and quantitative reliability, investments projects, and maintenance strategy, especially on developing new preventive methods for railroad bearing maintenance.

MARCELO NEVES received the B.Eng. degree in production engineering from UVV, in 2005, the specialization degree in maintenance engineering from ABRAMAN, in 2007, the specialization degree in rail cargo transport from the Military Engineering Institute, in 2009, and the specialization degree in maintenance management from UFPO, in 2013. He has been a Railway Engineer at VALE S.A., since 2005. He has experience in the areas of business risk management, basic maintenance guidelines, and investigation of railway occurrences, working in the areas of projects, railcar engineering, maintenance engineering, planning, scheduling and maintenance control, maintenance, management, and supervision of contracts, elaboration of operational standards and procedures, and development and optimization of spreadsheets and reports.