Unsupervised Probabilistic Anomaly Detection over Nominal Subsystem Events through a Hierarchical Variational Autoencoder

Alexandre Trilla[1,3], Nenad Mijatovic[2], and Xavier Vilasis-Cardona[3]

[1] *Alstom, Santa Perpètua de la Mogoda, Barcelona, 08130, Spain*
alexandre.trilla@alstomgroup.com

[2] *Alstom, Saint Ouen, Paris, 93482, France*
nenad.mijatovic@alstomgroup.com

[3] *DS4DS, La Salle, Universitat Ramon Llull, Barcelona, 08022, Spain*
xavier.vilasis@salle.url.edu

ABSTRACT

This work develops a versatile approach to discover anomalies in operational data for nominal (i.e., non-parametric) subsystem event signals using unsupervised Deep Learning techniques. Firstly, it builds a neural convolutional framework to extract both intrasubsystem and intersubsystem patterns. This is done by applying banks of voxel filters on the charted data. Secondly, it generalizes the learned embedded regularity of a Variational Autoencoder manifold by merging latent space-overlapping deviations with non-overlapping synthetic irregularities. Contingencies like novel data, model drift, etc., are therefore seamlessly managed by the proposed data-augmented approach. Finally, it creates a smooth diagnosis probabilistic function on the ensuing low-dimensional distributed representation. The resulting enhanced solution warrants analytically strong tools for a critical industrial environment. It also facilitates its hierarchical integrability, and provides visually interpretable insights of the degraded condition hazard to increase the confidence in its predictions. This strategy has been validated with eight pairwise-interrelated subsystems from high-speed trains. Its outcome also leads to further reliable explainability from a causal perspective.

1. INTRODUCTION

Anomalies are signs of a strange system condition that inherently represent a flaw, a degraded state, a fault, or a failure, and discovering them is of utmost importance to ensure the correct operation of a physical machine. The detection of anomalies using subsystem-event data is regarded as a traditional problem in the Prognostics and Health Management (PHM) community because it has a broad applicability but it still needs a definitive approach. This problem is assumed to be tractable using reams of data through a statistics-based perspective. However, there's no canonical approach to effectively process nominal events like these records. Specifically, there's a lack of consensus and methodology on algorithm selection in different scenarios (Huang, B., Di, Y., Jin, C., and Lee, J., 2017).

Subsystem event data are generally available through time-stamped nominal variables where typically no single message is decisive to raise an alarm. Thus, the density of information is low, along with the sparsity of this representation. These characteristics pose challenging encoding questions to the PHM engineers who are responsible for designing rules and procedures to diagnose anomalies in this environment. Such nominal event data have been commonly tackled as discrete-valued variables using counts of their occurrences in a sliding-time window, followed by a supervised learning scheme such as a Support Vector Machine or a Random Forest (Sammouri, W., Côme, E., Oukhellou, L., Aknin, P., and Fonlladosa, C.-E., 2014). After the Deep Learning revolution (Sejnowski, T. J., 2018), though, the recent state of the art in Anomaly Detection for PHM is dominated by the successive transformation of representations using Autoencoders, which are unsupervised neural networks that exploit the autoassociations in the data through a dense and efficient low-dimensional information-compressed embedded space (Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020).

Different solutions have been developed to address specific problems. For example, to counter the adverse effect of faulty data shortage and be robust to different operating conditions, an Extreme Learning Machine-based Autoencoder has been used to blend data from different sources conserving their

https://doi.org/10.36001/IJPHM.2023.v14i1.3431

homothety, and then its embedding has been used to classify the anomaly (Michau, G., and Fink, O., 2019). Similarly, for such open-set problems where the knowledge of all fault types may be incomplete at training time, the manifold of an adjusted Variational Autoencoders has been used (Arias Chao, M., Adey, B. T., and Fink, O., 2019). Also in this topology-preserving similarity line, further tweaks on the objective criteria to obtain a regular latent space have led to the consideration of Self-Organizing Maps within a Deep Autoencoder (Forest, F., Lebbah, M., Azzag, H., and Lacaille, J., 2019). Following this need for smooth behaviors, a recurrent Autoencoder has also been used to get continuous probabilities on machine health condition instead of the sudden evolution that is directly experienced when machines degrade (Shahid, N., and Ghosh, A., 2019). In light of all these approaches, it is clear that Autoencoders have generally been used with success as feature extractors and anomaly detectors for diverse applications (Farzad, A., and Gulliver, A., 2020; Dangut, M. D., Skaf, Z., and Jennions, I., 2020). Particularly, one of the most promising environments for this technique is found when the input data gets represented as an image and a convolutional Autoencoder architecture is deployed (Eid, A., Clerc, G., Mansouri, B., and Roux, S., 2021; Rodriguez Garcia, G., Michau, G., Ducoffe, M., Sen Gupta, J., and Fink, O., 2021).

This work unifies the former successful ideas under the same framework, and builds a novel value-added solution for maintainers to detect rolling-stock anomalies in a high-speed railway environment using *only operational data*. To this end, a generative approach is considered as its main component, being the most expressive probabilistic technique to model the complexity of the problem at hand. Moreover, this model naturally enables the production of synthetic data to face the shortage of anomalies that is typically found in a real-world commercial transport service. And finally, observing the industrial requirement of an interpretable safety-critical PHM system and its connection to visualization (Elattar, H. M., Elminir, H. K., and Riad, A. M., 2016), hazard maps are extracted to build trust with the customers and increase their confidence in this innovative approach.

The article is organized as follows: Section 2 describes the logged multi-subsystem operational event dataset and the framework to process it based on a Hierarchical Variational Autoencoder. Section 3 shows the diagnosis results obtained in terms of Anomaly Detection (i.e., a classification objective). Section 4 discusses the general interpretability insights that may be extracted, which are mostly based on causality, and Section 5 concludes the work with some future avenues of improvement.

2. MATERIALS AND METHOD

This section describes the data that have been used to learn and exploit the anomaly model, the strategy to obtain this knowledge, and the measurable key performance indicators to quantify the expected detection success in the field. Additionally, the ISO 13374 standard has been observed to design the proposed solution (ISO, 2003). What follows is a brief description of the main modules that have been implemented:

Data Acquisition The operational events have been logged using the Train Control Management System (TCMS), which is the on-board computer that sniffs the backbone network of the train.

Data Manipulation The subsystem event-data have been binarized into a logic-like waveform and arranged onto a charted geometric space.

State Detection The data-space has been transformed with filters and modeled using a probabilistic generative approach with latent variables. Additionally, synthetic data have been produced to enrich the model and generalize the diagnosis solution, which has been devised as a dichotomous classifier.

Advisory Generation Hazard maps have been produced to provide visual feedback of the degradation zones that are likely to generate anomalies.

2.1. Subsystem Event Dataset

While the trains are in commercial service, their on-board subsystems generate messages about their operation according to some predefined rules driven by specific events designed by their suppliers and manufacturers. These messages are then logged by the TCMS, which is continuously monitoring them. In this work, a dump of subsystem logs (syslogs) for a whole year has been collected from a high-speed rolling stock platform. What follows are some descriptive statistics of these records to better understand the nature of these longitudinal data.

The dataset amounts to 4.8M events distributed across the multiple train units in the fleet throughout the year, see Figure 1. There are two main modes in this distribution: trains that generated around 70k events, and trains that generated around 110k events. This may be due to different mission profiles to balance the load of the service.

These subsystem event data are essentially nominal, i.e., non-parametric. They are defined by a specific subsystem/train identification code and the timestamp of occurrence. Additionally, there are some context variables like the GPS location that may be useful to display operational details, and eventually to help fathom the potential reasons that may explain a given event pattern. For example, Figure 2 displays the evolution of monthly event counts showing seasonal patterns: this function is flat around 9k average unit events for half of the year, and plunges in the spring and the fall. Figure 3 displays the evolution of weekly events, showing that the service peak is on Thursday (busy business day) while the trough is on Sunday (late weekend). Finally, Figure 4 displays the event

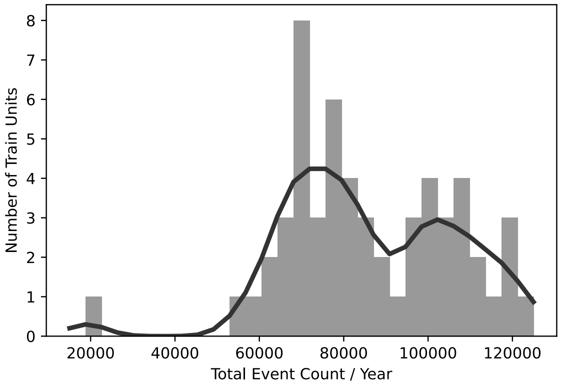

Figure 1. Histogram of the total event count per train unit, showing two main modes as humps in the kernel density estimation.

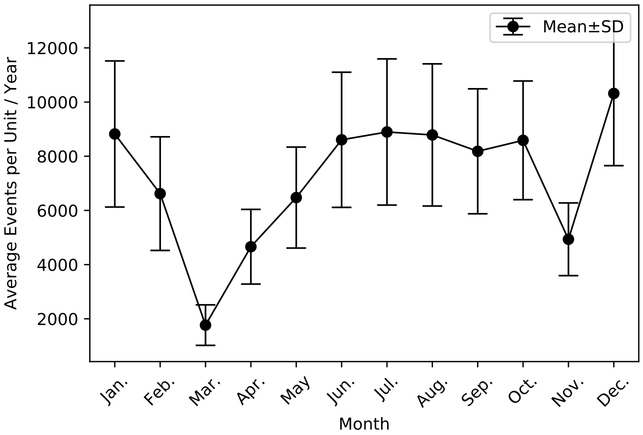

Figure 2. Monthly evolution of event counts.

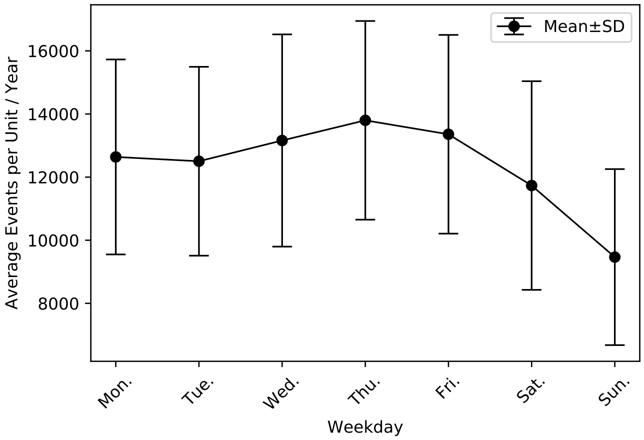

Figure 3. Weekly evolution of event counts.

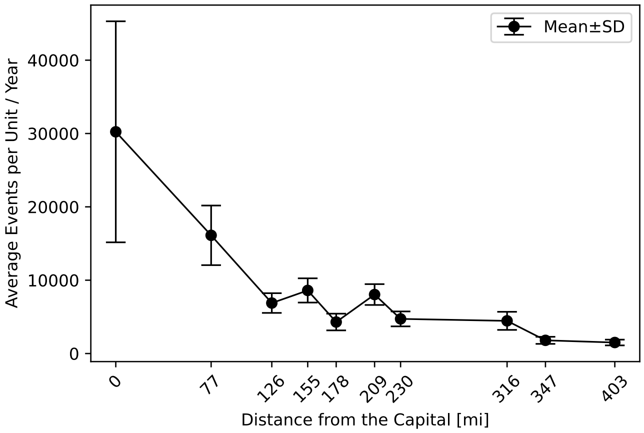

Figure 4. Evolution of event counts given the location.

evolution regarding the train location on the line, showing that the capital is the area where the majority of the events are generated, and the counts decrease exponentially on the more distant destinations.

Regarding the specific subsystems that issue messages into the network, Figure 5 displays their total arrangement. Additionally, for each of them, a power law defines its internal distribution of events, see Figure 6 for the Traction subsystem shown as an example. Note that there exists some functional spillover among the subsystems, for instance, between the Traction and the Brake. The rolling stock platform of use here equips a blended braking system by which the traction motor is both used to put the train into motion and also to stop it. This explains why braking events can be found in the Traction subsystem stream, e.g., “Traction/Brake Train Line Fault”, “Regenerative Brake Defect”, etc. This mixed nature of event occurrence justifies the importance of building a framework able to blend data from different sources. The next section describes how this point has been particularly considered in this research.

2.2. Anomaly Detection Framework

This section describes the solution that has been designed to detect anomalies in operational data using nominal subsystem events. Figure 7 shows its modular framework, where its functional blocks are shown in boldface, and the details of their implementation are further described in the following subsections.

2.2.1. Event-Voxel Data Fusion

In a PHM environment, the data that can reliably contain information about the failure of a machine is typically scarce. Therefore, all the data sources that may be within reach are advised to be collected and exploited, especially if a statistics-based approach is targeted (Gelman, A., 2021). However, the workload for data selection and filtering is significant with heterogeneous and complex datasets, especially in inference-

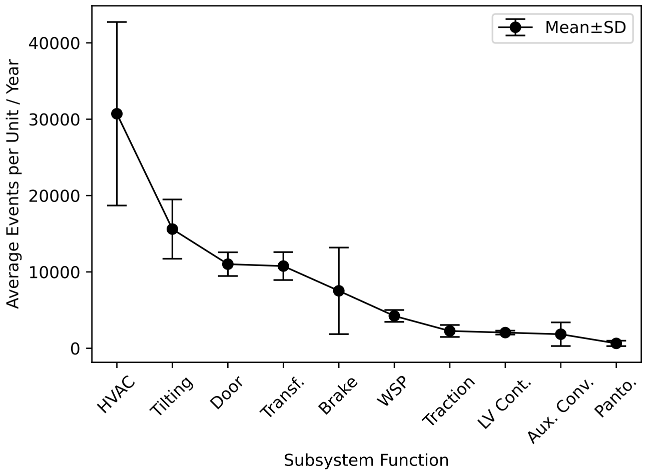

Figure 5. Ranked total event counts given the subsystems.

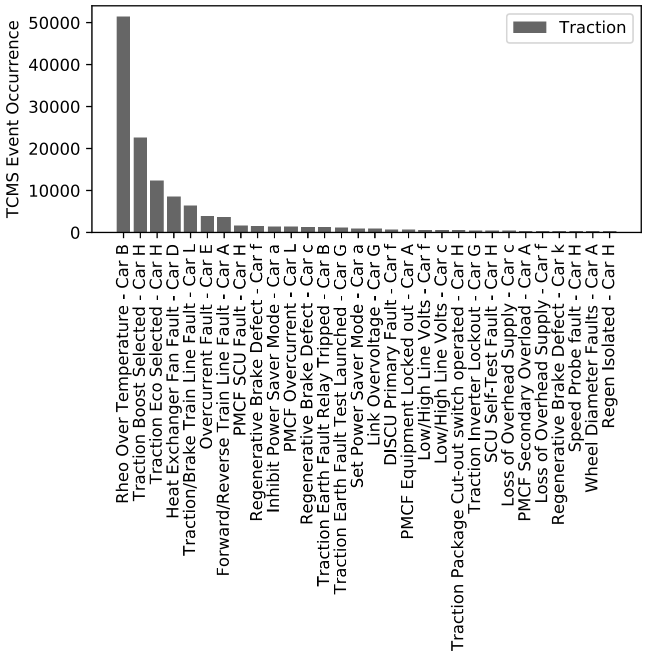

Figure 6. Histogram of the top 30 frequency-ranked events for the Traction subsystem.

based classification problems like Anomaly Detection (Huang, B., Di, Y., Jin, C., and Lee, J., 2017). In light of this scenario, there is a need to develop an automatic approach to represent and fuse different data from distinct origins (Hu, X., Eklund, N., and Goebel, K., 2007), i.e., concurrent intrasubsystem as well as intersubsystem sources. The proposed process to attain this goal is described as follows.

Initially, the data from the timestamped subsystem events are massively processed using regular expressions to extract the key-value pairs and conflate similar logs into matching clusters (Du, M., Li, F., Zheng, G., and Srikumar, V., 2017). Additionally, they are segmented into train units and 24-hour time sets that align with the commercial transport schedule, yielding around 20k instances within the dataset. Also, the coordination with the maintenance activities runs at the day-by-day level, thus the decisions are made by the Operations Team within this time frame. Finally, the resulting sets undergo the subsequent series of dimensional (D) transformations:

1D: Nominal Event to Parametric Time Series The nature of the nominal event data is first transformed into a time series of binary parametric variables using a spreading filter (Hu, X., Eklund, N., and Goebel, K., 2007). The resulting time-dilated data resemble the pulse signals of a logic circuit that can be further analyzed because they represent useful information for health management such as the time between events (Xie, Y. J., Tsui, K. L., Xie, M., and Goh, T. N., 2010). The resolution in time adopted in this work is of 30 minutes, i.e., 48 time slices per day.

2D: Intrasubsystem Diversity To illustrate the information that a single subsystem generates by itself, e.g., see Figure 6, a bidimensional image-like representation is proposed. Such charted data organization can display complex patterns such as correlations, recursive behaviors, or spectral components (Rodriguez Garcia, G., Michau, G., Ducoffe, M., Sen Gupta, J., and Fink, O., 2021; Eid, A., Clerc, G., Mansouri, B., and Roux, S., 2021). In this work, the 30 most frequent events per subsystem are considered. To see how this representation is effective to display different degradation conditions, Figure 8 shows a Normal instance chart of Traction subsystem behavior. In this representation, only the most frequent events at the top of the rank get generated sparsely. In contrast, Figure 9 shows an Anomaly instance chart. In this case, many events get generated concurrently, also in the infrequent event space. These two plots show the two extremes of the degradation spectrum. For predictive maintenance purposes, the interesting analysis lies in the transition phase, especially around the incipient point of failure.

3D: Intersubsystem Diversity The last step in the representation of the multiple subsystem data adds a new dimension where different charts may be stacked. This approach clearly shows the concurrent nature of event observation among the different generators. In this work, pairwise-interrelated subsystems such as the Traction and Brake example are considered.

In the proposed volumetric representation, the smallest quantum of data is therefore given by a voxel of time, intrasubsystem and intersubsystem binary event occurrence. These voxels are then arranged into a tensor of size (30,48,2) that is suitable for exploitation with a Deep Learning model, as is described in the next section, to extract the relevant dynamic (i.e., time evolving) data characteristics between the thirty most frequent

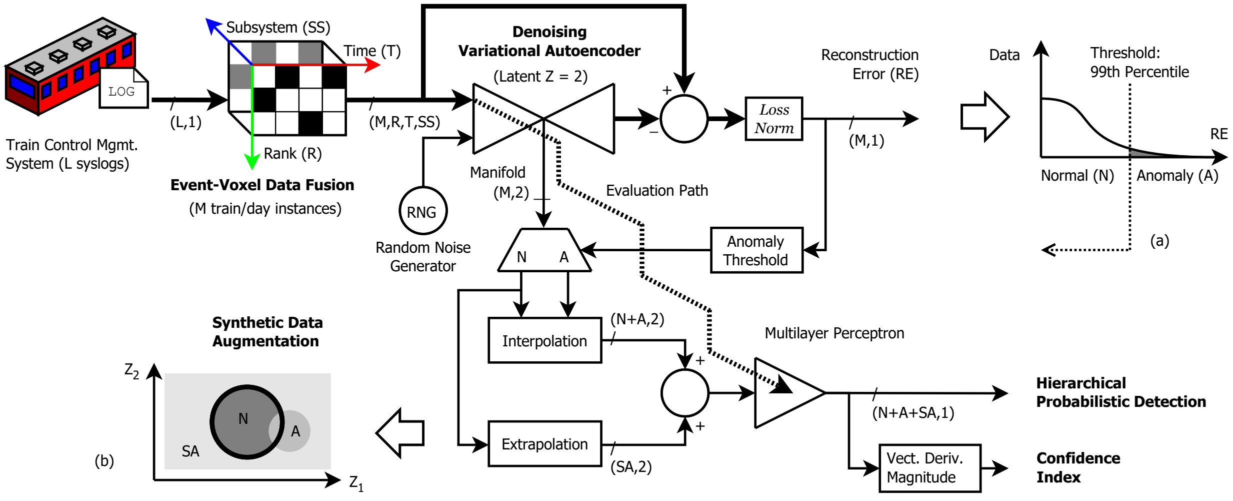

Figure 7. Diagram of the proposed Anomaly Detection framework. Plot (a) depicts the expected distribution of the Reconstruction Error. Plot (b) depicts the expected representation on the augmented latent space. This design is mostly focused on training the solution. Regarding its industrial deployment, the data path for its straightforward diagnosis evaluation is displayed as a thick dashed line connecting the manifold in the Variational Autoencoder with the Multilayer Perceptron to estimate the probability of anomaly.

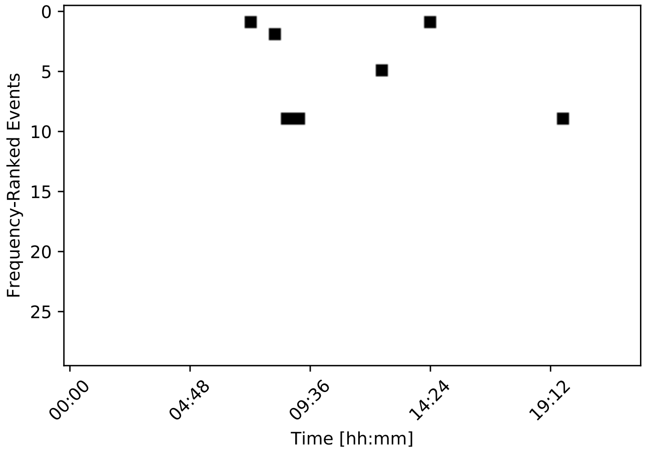

Figure 8. Chart representation of a Normal condition pattern.

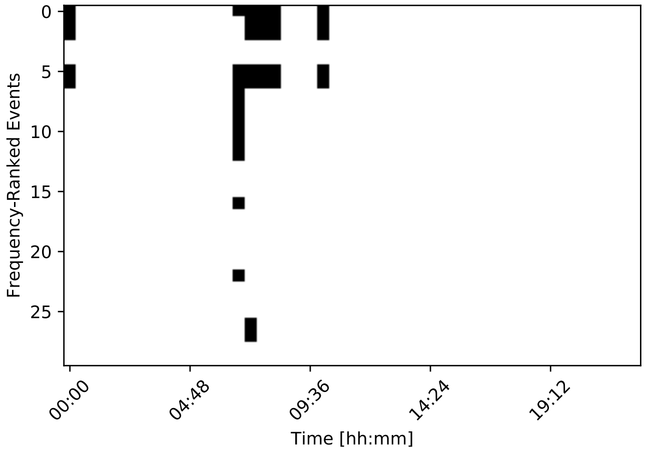

Figure 9. Chart representation of an Anomaly condition pattern.

events for two related subsystems (e.g., the Traction and the Brake).

2.2.2. Denoising Variational Autoencoder

A Variational Autoencoder (VAE) is a probabilistic approach that is used to represent the process of data generation. The VAE provides a principled framework for learning deep latent-variable encoding models $Q(z)$, and the corresponding decoding inference models (Kingma, D. P., and Welling, M., 2019). This method is a key enabler to implement the proposed integrated approach working on unsupervised categorical data X like regular operational events (Hancock, J. T., and Khoshgoftaar, T. M., 2020). At its core, the VAE is a variational Bayesian method (Doersch, C., 2016), and given that the Bayesian theory rests on an axiomatic foundation, the VAE is guaranteed to have quantitative coherence that other methods do not have (Duda, R. O., Hart, P. E., and Stork, D. G., 2001). Moreover, adding random noise and regarding a denoising learning schedule is helpful to secure a good generalization performance of the model and enable its reuse for pretraining on downstream tasks (Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., and Vincent, P., 2009).

The VAE fundamentally maximizes the probability of the data under the entire generative process, i.e., through the compression in the embedded latent space. Its objective function is the Evidence Lower Bound (ELBO), see Eq. (1), where KL is the Kullback-Leibler divergence. The three main factors that define the implementation of the ELBO for the proposed Denoising VAE are listed as follows:

- Encoding/Decoding Functions Q: Convolutional Neural Networks
- Latent Space Manifold z: Multivariate Normal Distribution
- Reconstruction Error/Loss: Binary Cross-Entropy

$$\begin{aligned} ELBO(X,Q) =& E_{z\sim Q}[\log P(X|z)] - KL[Q(z)\|P(z|X)] \\ =& E_{z\sim Q}[\log P(X|z)] - \\ & E_{z\sim Q}[\log Q(z) - \log P(z|X)] \end{aligned} \tag{1}$$

Regarding the encoding, the representation of the nominal event data X into 3D binary voxels arranged into tensors naturally leads to their effective exploitation through a deep convolutional neural framework. Expressive complex functions in Q are to be learned with the embedded non-linearities, which are introduced by the Rectified Linear Unit (ReLU) activation function, and the weight-sharing strategy of its filters help the resulting network to not overfit the data. Moreover, events are well-aligned at similar scales, which results in less variation in the critical data (Kanazawa, A., Sharma, A., and Jacobs, D., 2014). Finally, introducing random noise at this stage (e.g., through a few voxel value flips) plays an important role in achieving good generalization performance: it makes nearby data points in the low dimensional manifold robust against the presence of small deviations in the high dimensional observation space (Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A., 2008). This variation could be physically interpreted as the thermal noise in the sensors that eventually generate the events in the subsystems.

Regarding the learned embedding, each dimension of the latent random variable z is assumed to be independent of each other (i.e., they are factorized) and modeled by a univariate Gaussian distribution whose parameters (i.e., the mean and the variance) are obtained by the non-linear neural encoding function Q. As a result, the latent space displays enough smooth regularity to be considered as a manifold. Specifically, a manifold is a topological space that is locally Euclidean (Bredon, G. E., 1995). This low-dimensional geometric analysis makes it computationally advantageous compared to the high dimensional input. Additionally, this latent distributed representation, which is set to 2 dimensions for representational purposes, is amenable to the visual interpretation of the hazardous anomaly zones. This is extremely useful because the similarity in high dimensional spaces is meaningless (Fefferman, C., Mitter, S., and Narayanan, H., 2016). Moreover, limiting the expressiveness of this bottleneck layer helps to compress the data and thus retain its most meaningful attributes, which is likely to be helpful for the generalization of the solution and prevent overfitting. Finally, given that stochasticity is inherent in the sampling process on the manifold (here this can be taken for a sort of injected latent noise), further improved performance is expected (Im, D. J., Ahn, S., Memisevic, R., and Bengio, Y., 2017). The source of this variation could be physically found in the seed of the random number generator, e.g., a timer.

Regarding the objective loss function, most PHM approaches dealing with parametric data assume Gaussian or Laplacian error likelihood distributions and thus consider Mean Squared or Mean Absolute Error (MAE) metrics to train and evaluate their performance (Rodriguez Garcia, G., Michau, G., Ducoffe, M., Sen Gupta, J., and Fink, O., 2021). MAE is especially robust to outliers in time series data (Lai, G., Chang, W.-C., Yang, Y., and Liu, H., 2018), thus helping in the modeling of the regular operational condition. Nevertheless, for the current event-based scenario, interpreting binary data as probabilistic targets and introducing classification metrics such as the Binary Cross Entropy leads to faster training as well as improved generalization (Simard, P. Y., Steinkraus, D., and Platt, J. C., 2003). This implicitly assumes that the reconstruction error in the ELBO is Bernoulli distributed (Sicks, R., Korn, R., and Schwaar, S., 2020).

Finally, to complete the description of the VAE proposed in this work, Table 1 shows some further details about the internal structure and parameters for the Encoder part (note that the Decoder simply mirrors and unwinds this given configuration). In total, the VAE comprises over 120k trainable parameters.

2.2.3. Synthetic Data Augmentation

To enhance the out-of-distribution generalizability and the robustness of the proposed solution, the available data is augmented. This gives rise to a set of synthetic instances that are expected to go beyond the limited set of observed anomalies. This strategy is increasingly gaining adoption in the industry (Strickland, E., 2022), where the assets are typically overmaintained to minimize the risk of a service-affecting failure.

In the previous section, the management of noise was described (along with the introduction of a denoising strategy) for performance improvement purposes (Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A., 2010). Additionally, the data is here transformed by considering shifts in time, also known as translations. Convolutional Neural Networks are not naturally invariant to translations, but they can

Table 1. VAE Encoder structure parameter chart.

Layer Name	Type	Filter	Stride	Amount	Activation	Output Shape	Parameters
Event Voxel	Input				Linear	(30, 48, 2)	0
Shallow Receptive	Conv2D	(3,3)	2	32	ReLU	(15, 24, 32)	608
Deep Receptive	Conv2D	(3,3)	3	64	ReLU	(5, 8, 64)	18496
Sparse Vector	Flatten					(2560)	0
Dense Vector	Dense				ReLU	(16)	40976
Latent Mean	Dense				Linear	(2)	34
Latent Variance	Dense				Linear	(2)	34

acquire this feature if such transformation is embedded in the data strategy (Biscione, V., and Bowers, J. S., 2021), especially when no Pooling layers are introduced in the pipeline (Chaman, A., and Dokmanic, I., 2021), as is the case here. Eventually, the data are separated into Normal and Anomaly groups according to their amount of reconstruction error, which is a reliable indicator to detect anomalies when its value is over the 99th percentile (Rodriguez Garcia, G., Michau, G., Ducoffe, M., Sen Gupta, J., and Fink, O., 2021). What follows is the description of the synthetic generation process based on interpolation and extrapolation driven by this anomalous condition distinction, all of which take place in the latent space manifold that has been designed to exhibit enough regularity to perform these operations.

The few instances that are regarded as anomalous, i.e., the ones that display a large reconstruction error, comprise the minority class as they lie on the long tail of the loss distribution. This data imbalance can cause learning problems and result in skewed outcomes. To counter this adverse situation, a combination of oversampling for the minority (i.e., Anomaly) class and undersampling for the majority (i.e., Normal) class achieves better classifier performance (Chawla, N. V., and Bowyer, K. W., 2002). Specifically, the method for oversampling the minority class involves linearly interpolating among the nearest neighbors, which thus creates similar synthetic examples.

Finally, generative models like the VAE give rise to “fantasy” data whose probability distribution is the same as that of the observed data (Bishop, C. M., 2006). This principle is exploited here outside the main cluster of Normal data as a grid of non-overlapping instances deployed on the latent space (Huh, D., 2011). In PHM, particularly, this extrapolation-based approach was originally inspired by the natural immune system (Qiu, H., Eklund, N., Hu, X., Yan, W., and Iyer, N., 2008), and thus there is sensible evidence to believe in its effectiveness.

2.2.4. Hierarchical Probabilistic Detection

Beyond the plain discriminative function introduced by the amount of reconstruction error, providing a fine-grained assessment of the stage of degradation is advantageous to avoid a sudden evolution from Normal to Anomaly conditions (Shahid, N., and Ghosh, A., 2019). To this end, a Multilayer Perceptron (MLP) neural network is hierarchically introduced on the manifold z to directly estimate the probability of Anomaly p_A, see Eq. (2) for a matrix notation of this classification function, where W are the input (I) and hidden (H) transformation matrices, and g is a non-linearity bounded between 0 and 1 such as the logistic sigmoid function. The computed probability enables considering decision theory criteria such as the management of risk driven by the reject option, and also facilitates its combination within more integrated probabilistic solutions (Bishop, C. M., 2006).

$$p_A(z) = g(W_H(g(W_I z))) \tag{2}$$

Well-regularized MLP’s significantly outperform recent state-of-the-art specialized architectures (Kadra, A., Lindauer, M., Hutter, F., and Grabocka, J., 2021). Functionally, the MLP performs a non-linear logistic regression that learns the tessellation of the latent space and decouples the two degradation conditions. This objective is attained by the contrastive character of the cross-entropy loss (Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D., 2020), which is fueled by the thresholded reconstruction error that is incorporated explicitly as a binary target within a supervised training process (Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M., 2014).

2.2.5. Confidence Index

To close the design of an industrial system, indicating the amount of trust in the system’s outcome is useful for the consumer of this information. This goal is related to the estimation of the uncertainty in the given solution. In this paper, the smoothness of the probabilistic anomaly detection function p_A is exploited as follows: the Confidence Index (CI) is ultimately described by the rate of its change. This inherently implies that the transition zones are unstable and uncertain, while the plateaus are stable and certain. Given that the detection function depends on the distributed representation of the bidimensional manifold z (that is locally Euclidean), the magnitude of its vector derivative $\nabla = (\partial/\partial z_1, \partial/\partial z_2)$ is what is taken for reference to indicate confidence in the prediction, see Eq. (3). Finally, a unitary bound on the resulting CI is introduced for normalized advisory purposes.

$$CI(z) = 1.0 - \min(\|\nabla p_A(z)\|, 1.0) \tag{3}$$

2.3. Performance Evaluation

In most real-world settings, the probability of an anomaly is expected to be only slightly greater than zero (Wu, R., and Keogh, E., 2021). In this sense, the purpose of this section is to validate that the proposed probabilistic approach effectively *models* the degradation of the rolling stock using nominal subsystem events. As a result, the probability of Anomaly must be strictly higher for the degraded condition than for the Normal (i.e., regular) condition. To do so, a balanced sample of validation data is obtained after the discrimination determined by the amount of reconstruction error, see Section 2.2.3. 10% of the anomalous instances are included in this hold-out validation sample, which amounts to 120 examples in total.

The key performance indicators for this evaluation are driven by the probability of Anomaly p_A for both the Normal and the Anomaly evaluation sample. Gaussianity in the distributions is assumed for statistical convenience, because the probability is a bounded quantity between 0 and 1. Also, the customary minimum of 30 instances to reliably estimate the two statistical moments of this distribution type (i.e., the mean and the variance) are guaranteed in the evaluation sample (Lejeune, M., 2010). The significance of their mean average differences is determined by the Student's t-test (Gosset, W. S., 1908). Further classification evaluation can be easily attained by introducing a threshold to discretize the probabilistic decision, which may also help to manage the potential reject option. The specific value of this threshold is typically set at 0.5, i.e., in the middle of its range. The Precision P and Recall R measures that succeed consider the impact of False Positive FP and False Negative FN errors respectively with regards to the True Positive TP successes, which are all to be found in the confusion matrix, see Eq. (4).

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \tag{4}$$

Finally, the limitations of the proposed VAE-based Anomaly Detection approach define the epistemic uncertainty in the model. To determine the range of their impact on the diagnosis performance, the following evaluation environments are considered (for practical experimental purposes, only the subsystems that generate most of the events are taken into consideration in this work):

- Locomotion: Traction + Brake
- Indoors: Heating, Ventilation, and Air Conditioning (HVAC) + Doors
- Bogie: Tilting System + Wheel Slip Protection (WSP)
- Energy: Transformer (Transf.) + Auxiliary Converter (Aux. Conv.)

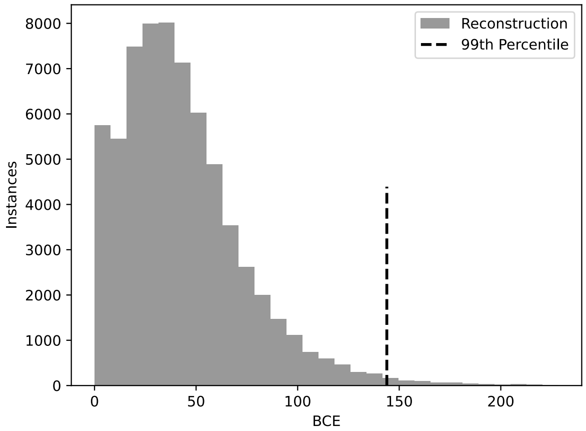

Figure 10. Histogram of the Binary Cross Entropy (BCE) Reconstruction Error along with the 99th percentile threshold. The plot roughly matches the expected distribution of this Loss, see Figure 7(a).

3. RESULTS

This section presents the results obtained with the proposed Anomaly Detection approach based on operational subsystem event data. Figure 10 shows an example of the the distribution of degradation provided by the histogram of the Reconstruction Error/Loss. The mass of this distribution is largely skewed toward the lower end, and it decays exponentially as the instances become increasingly anomalous (this is the expected behavior at the fleet level). A statistical threshold over the 99th percentile is used to separate the Normal from the Anomaly conditions. This criterion works well in the real world to spot actual anomalies (Rodriguez Garcia, G., Michau, G., Ducoffe, M., Sen Gupta, J., and Fink, O., 2021). Moreover, on this distribution there seem to be two modes of behavior, a small one that aligns with the zero origin, and a large one that is somewhat shifted. This may be associated with the different regimes of the trains, e.g., low-speed maneuvering close to the depot/station (i.e., the low volume of records) and high-speed intercity transit (i.e., the majority of the records).

Delving deep into the internal operation of the system, Figure 11 shows the tessellation of the bidimensional latent manifold. In this hazard map, the decision boundary (i.e., $p_A = 0.5$) wraps the instances that are deemed to be Normal, and leaves out the ones that belong to the Anomaly category or the synthetic outliers. Additionally, Figure 12 displays the confidence in the diagnostic, which essentially depicts the silhouette of the Normal region. As expected, the transition zone is the most uncertain point.

Finally, Table 2 shows the performance of the Anomaly Detection approach for each of the evaluation environments. In all cases, the average probability of abnormality for the Anomaly

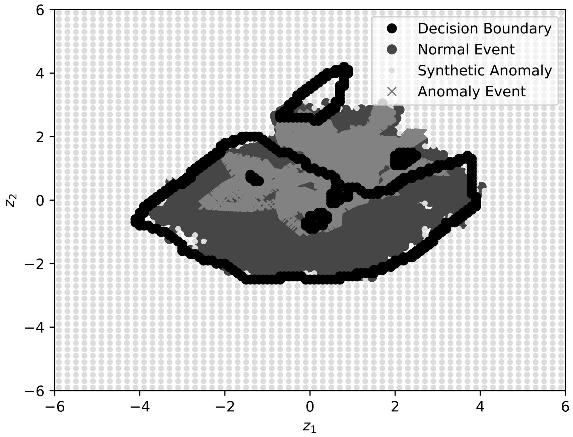

Figure 11. Tessellation of the latent manifold on the learned bidimensional embedding $z = (z_1, z_2)$. The probabilistic anomaly decision boundary is shown at $p_A(z) = 0.5$, which is the random guess on a dichotomic classification problem. Note that while the latent space is continuous, the evaluation points are necessarily discrete, and a visually dense grid has been used here to display the Normal closed region. While a continuous function approximating this boundary is likely to be faithful to reality, only the spots that have been actually evaluated are represented. The plot matches the expected distribution of this embedded space, see Figure 7(b).

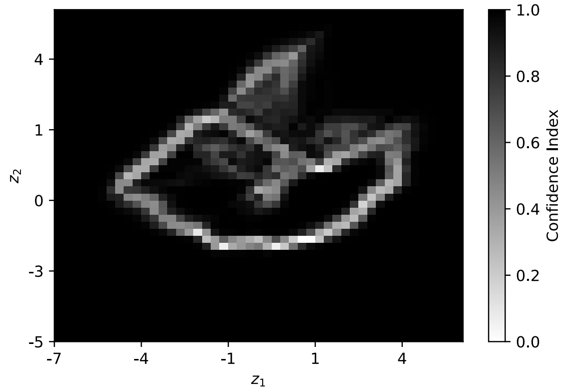

Figure 12. Confidence Index shown on the latent manifold related to Figure 11.

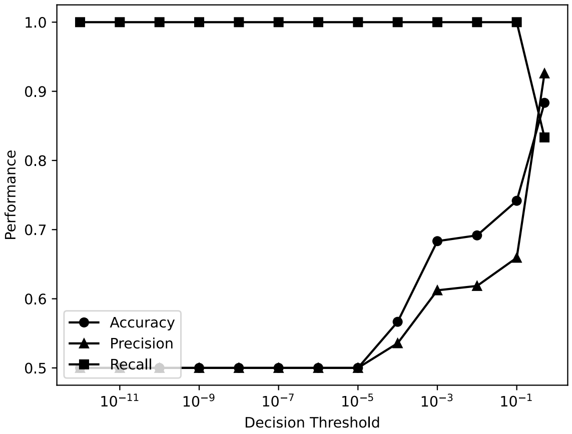

Figure 13. Precision and Recall curves driven by the sensitivity of the Decision Threshold. Accuracy is also shown here only for reference as the total rate of correct classifications.

condition is significantly greater than for the Normal regular case. The resulting range of classification performance indicators lies around 80%, which is similar to a historical baseline obtained on comparable data (Sammouri, W., Côme, E., Oukhellou, L., Aknin, P., and Fonlladosa, C.-E., 2014). See Figure 13 for the impact of the decision criterion on the types of error displayed by the system. A smaller threshold value drives the system toward conservatism (i.e., high Recall at the expense of false alarms), while a greater value yields an eager behavior (i.e., high Precision at the risk of missing a failure).

4. DISCUSSION

This section addresses some typical qualms about time-series based anomaly detection, and provides insights into its interpretability from a causal perspective.

4.1. Reliability

Conventional performance indicators for anomaly detection methods based on time-series data can sometimes be misleading (Wu, R., and Keogh, E., 2021). This happens, for example, when the signals are so trivial that a single descriptive statistic such as the mean or the standard deviation suffices to explain them, or where the anomalies are directly found at the end of the data sequence (e.g., on run-to-failure tests). None of these situations apply to the scenario tackled in this work. In hindsight, though, simplifications to the proposed approach could now be found, but these seem unlikely to be have been devised initially with the data only.

Perhaps one aspect worth discussing here is the noise in the labels, which is a pervasive problem in the field because manual expert-labeling of each instance at a large scale is not feasible (Kim, S., Choi, K., Choi, H.-S., Lee, B., and Yoon, S., 2022). This work, albeit framed in an unsupervised learning setting, relies on the signal reconstruction error as an *imperfect surrogate* for the ground truth, which is used to estimate the probability of Anomaly with the cross-entropy loss. Here, the 99th percentile loss drives this discriminative labeling criterion, motivated by its reported success to identify anomalies in the real world (Rodriguez Garcia, G., Michau, G., Ducoffe, M., Sen Gupta, J., and Fink, O., 2021). However, if this high value is reduced, the detection results are likely to be differ-

Table 2. Detection performance driven by the probability of Anomaly, that is applied to the Normal (N) and Anomaly (A) validation instances, taking into account their environments. Statistical mean μ and standard deviation σ are computed, along with the *p*-value of the significance *t*-test, and the Precision/Recall values at the decision boundary of $p_A = 0.5$.

Environment	$\mathbf{p_A(N)}[\mu/\sigma]$	$\mathbf{p_A(A)}[\mu/\sigma]$	***p*-value**	**Precision**	**Recall**
Locomotion	0.18/0.19	0.78/0.25	6e-28	0.92	0.83
Indoors	0.21/0.29	0.70/0.28	1e-15	0.82	0.71
Bogie	0.39/0.18	0.66/0.25	7e-10	0.72	0.60
Energy	0.23/0.20	0.76/0.34	7e-18	0.91	0.72

ent, perhaps affecting the capacity of the system to deal with instances increasingly similar to regular data.

In such a hybrid learning environment, if the training data is “corrupted” with this pseudo-label, deep models such as the VAE tend to overfit the noise, thereby achieving poor generalization performance (Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B., 2020). This effect can be observed as a condition overlap in the latent space, see Figure 11, although this region also shows a lower Confidence Index, see Figure 12. Moreover, this Bernoulli-distributed error makes it difficult to identify out-of-distribution instances when there are lots of zeroes in the data (Yong, B. X., Pearce, T., and Brintrup, A., 2020), as is the case with the sparse subsystem events, see Figures 8 and 9. Nevertheless, when the ReLU is the only non-linearity in the system (check Table 1), the loss curvature is immune to class-dependent label noise (Patrini, G., Rozza, A., Menon, A., Nock, R., and Qu, L., 2017), which increases the confidence in the proposed approach.

4.2. Causal Explainability

Section 2.1 briefly described the blended braking system and the impact that one subsystem has on another, i.e., Brake on Traction. The Locomotion environment is very illustrative and further interesting insights may be extracted. This section is dedicated to providing such explanations, especially form the perspective of the inferred causality (Zaman, N., Apostolou, E., Li, Y., and Oister, K., 2022).

Causal inference is here motivated by the Kullback-Leibler divergence, which is used in the objective function of the VAE, see Section 2.2.2. It turns out that this value is a suitable measure of causal influence (Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schölkopf, B., 2013). Therefore, the question naturally arises: has the VAE automatically learned any cause-effect relationships?

4.2.1. Graphical Causal Structure

In this work each dimension of the latent space is assumed to be an independent Gaussian, see Section 2.2.2 for further details. This design choice creates a disentangled representation that is not necessarily causal, it has been introduced only to allow a more complex joint distribution to be constructed from simpler components (Bishop, C. M., 2006). To progress toward a semantically interpretable system, *causally* disentangled latent variables are needed. These can in fact be obtained from VAE models using an embedded layer to transform independent exogenous factors (i.e., the root causes) into causal endogenous ones (i.e., their effects) that correspond to causally related concepts in the data (Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J., 2020). However, the data must already contain sample-wise causal labels to learn this richer representation. In the absence of such cues, this section uses a Causal Discovery approach to create a potential graphical description of the inherent causal structure.

Considering that the available event subsystem data can be framed as a multivariate time-series of binary variables, causality is expected to be observed as the precedence of events. To capture their causal links, the Peter-Clark (PC) algorithm is proposed (Spirtes, P., Glymour, C., and Scheines, R., 2001). PC is a causal network learning algorithm that copes well with high dimensionality and can often also identify the direction of contemporaneous links (Runge, J., Bathiany, S., Bollt, E. *et al.*, 2019). It is one of the oldest algorithms that is consistent under i.i.d. sampling assuming no latent confounders, i.e., all relevant variables need to be observed in the data (Glymour, C., Zhang, K., and Spirtes, P., 2019). The PC algorithm starts by building a fully-meshed graph with all the variables, and then evaluates the strength of the associations by testing their conditional independence using the time-series data. Eventually, it removes those edges that display zero partial correlation. Finally, it applies a series of heuristics to orient the links that remain giving them a causal direction, and the resulting graphical structure is provided.

In this analysis, the top 10 frequency-ranked events are considered, 5 for each subsystem in the Locomotion environment, see Table 3. Event simultaneity is expected, especially in the presence of anomalies. Figure 14 shows the generated causal graphical structure.

Based on these results, the subsystem interrelation between the Brake and the Traction is mostly evident, e.g., rheostat over temperature (T1) is caused by a failure on the blended braking system (B4 and B5) and on the fan of the heat exchanger (T4). In some cases, though, these associations are not so clear-cut. For example, the 5th Traction event (i.e., T5), which specifically refers to a “Traction/Brake fault”, is not caused by any of the most frequent Brake events according to the criteria

Table 3. Description of the top-ranked subsystem events in the Locomotion environment.

Event Rank	Traction (T)	Brake (B)
1	Rheo Over Temperature	Brake Supply Pressure High
2	Traction Boost Selected	Parking Brake Applied Pressure Switch
3	Traction Eco Selected	Main Line Pressure High
4	Heat Exchanger Fan Fault	Application Error 1 (blending)
5	Traction/Brake Train Line Fault	PWM Signal 2 Dyn Brake Out of range

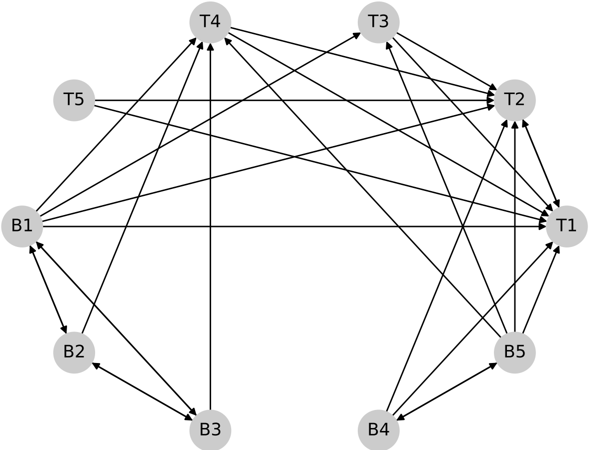

Figure 14. Causal graph for the Locomotion environment, i.e., including the Traction (T) and Brake (B) subsystems. Node name code: {Subsystem}{Rank}. See Table 3 for further details. Arrows indicate event association from cause to effect.

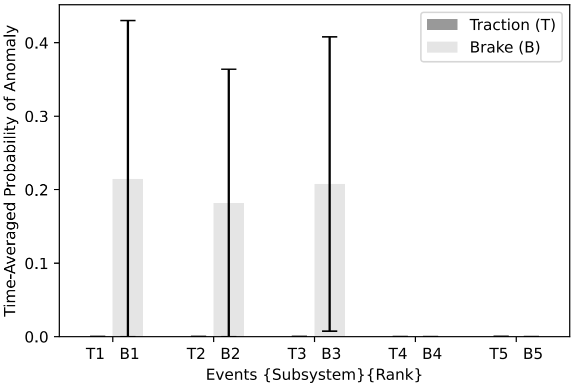

Figure 15. Sensitivity analysis on the Locomotion environment. See Table 3 for further details. Assuming Normality for the day-level average distributions, bar heights indicate their mean values, and whiskers indicate one standard deviation. All the visually imperceptible bars actually have a negligible probability in the order of 10^{-4}.

of the PC Causal Discovery algorithm.

What is more, the graph shows some bidirected edges, e.g., among B1, B2, and B3. This is likely to indicate the presence of an unobserved confounder, which reveals a limitation of the PC approach: since its outcome is a Markov equivalence class, there is likely to be another (possibly better) graphical representation that explains the same data. In fact, direct PC application is not advised for the time series case, despite its apparently good results, and other more involved methods using more powerful statistical tests with time lags should be explored on top of it (Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D., 2019). Additionally, the subject matter experts should elucidate these effects and resolve the causal directionality conflict. However, the PC algorithm serves well to make the point of the discussion, and its result constitutes a solid basis for further research.

4.2.2. Sensitivity Analysis

In the context of this work, the sensitivity analysis of interest determines how the probability of Anomaly is affected by changes in the subsystem event data. This may help quantify the maximum bias that is reasonably expected for unmeasured confounding (Hernán, M. A., and Robins, J. M., 2020), which was detected by the former Causal Discovery approach (also note that the VAE model implicitly assumed that the events are independent). Here, a time-averaged analysis at the day level of the top-ranking Locomotion events is performed, see Figure 15.

This sensitivity study shows that the impact of the Traction is barely noticeable compared to the impact of the Brake, especially regarding its three most frequent events, which are also the ones subject to an unobserved confounder. Taking all this extracted information into account, it could be stated that whenever an anomalous situation occurs and a Traction event is generated, the actual root cause is likely to be found on the Brake. However, causality at the model level cannot be extrapolated to the real world (Molnar, C., 2019). It is a global interpretation of the available observational (i.e., ambiguous) data. Unless further expert criteria are additionally considered, these results may ultimately be driven by correlation, as this point cannot yet be fully rejected. The contrapositive argument that no-correlation implies no-causation could explain some of these results, especially for the 4th and 5th event ranks, which display a null risk of Anomaly. In the end, both correlation and convolution are linear shift-invariant operators (Szeliski, R., 2022), and since the latter defines the structure of the VAE, it could also help elucidate this behavior.

5. CONCLUSION

The strategy to detect anomalies using only operational data through a Hierarchical Variational Autoencoder has provided good results on par with previous experience. Moreover, the fine-grained probabilistic diagnosis has enabled 1) tackling the gradual degradation process that is observed at the fleet level, 2) building interpretable visual insights through hazard maps, and 3) assessing the confidence in the predictions.

Although the focus of the paper is on subsystem event streams as a challenging signal source, the method can be readily transferred to other domains (including other types of trains) using parametric data typically used in PHM: the convolutional structure can be directly applied to vibration, current, pictures, etc. What is more, all these environments may be ultimately merged into an ensemble towards a complete holistic solution where, for instance, the events of the Brake subsystem could be complemented with the shudder of a brake disk (e.g., from an accelerometer) and the thickness of the brake pads (e.g., from a camera).

This work has relied mainly on the management of random noise as a means to increase the robustness of the solution. However, interesting improvement directions may be devised when considering alternative loss functions in the VAE that are robust to outliers such as the Tsallis entropy (Sârbu, S., and Malagò, L., 2019), the coupled entropy (Cao, S., Li, J., Nelson, K.P., and Kon, M.A., 2022), the tamed cross-entropy (Martinez, M., and Stiefelhagen, R., 2018), and the hyperbolic cosine loss (Chen, P., Chen, G., and Zhang, S., 2019).

Moreover, this work has focused on providing a probabilistic function for the degradation of the assets, and the confidence in its outcome has been resolved using the magnitude of its gradient. Perhaps it could be more reliable to quantify the uncertainty (i.e., the variability) in the prediction using dropout in the MLP or introducing some fluctuations in its input latent representation, thus keeping a probabilistic description of the confidence. This is regarded as interesting future work.

Finally, the representation of causality is also a topic that deserves further attention (Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y., 2021). The Discussion has already revealed some straightforward insights, but a deeper understanding is necessary to make stronger conclusions. This paves the way for the consideration of Deep Learning to directly manage the construction of a Structural Causal Model from first principles (Zečević, M., Dhami, D. S., Veličković, P., and Kersting, K., 2021), and be able to identify the cause-effect relationships that describe the degradation processes in full detail.

ACKNOWLEDGMENT

We would like to show our gratitude to our colleagues Dr. Jonathan Brown and Quentin Possamaï for their insightful comments which greatly improved the manuscript. The contribution of Alexandre Trilla to this research was partially supported by the Government of Catalonia (Generalitat de Catalunya) Grant No. 2020 DI 54.

REFERENCES

Arias Chao, M., Adey, B. T., and Fink, O. (2019). Knowledge-Induced Learning with Adaptive Sampling Variational Autoencoders for Open Set Fault Diagnostics. *arXiv:1912.12502 [cs.LG]*, 1–21.

Biscione, V., and Bowers, J. S. (2021). Convolutional Neural Networks Are Not Invariant to Translation, but They Can Learn to Be. *Journal of Machine Learning Research*, *22*(229), 1–28.

Bishop, C. M. (Ed.). (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc.

Bredon, G. E. (1995). *Topology & Geometry*. Springer-Verlag.

Cao, S., Li, J., Nelson, K.P., and Kon, M.A. (2022). Coupled VAE: Improved Accuracy and Robustness of a Variational Autoencoder. *Entropy*, *24*(423), 1–25.

Chaman, A., and Dokmanic, I. (2021). Truly shift-invariant convolutional neural networks. *Proc. of the IEEE / CVF Computer Vision and Pattern Recognition Conference*, 3773–3783.

Chawla, N. V., and Bowyer, K. W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, P., Chen, G., and Zhang, S. (2019). Log Hyperbolic Cosine Loss Improves Variational Auto-Encoder. *Proc. of the International Conference on Learning Representations*, 1–15.

Dangut, M. D., Skaf, Z., and Jennions, I. (2020). Rare Failure Prediction Using an Integrated Auto-encoder and Bidirectional Gated Recurrent Unit Network. *IFAC PapersOnLine*, *53*(3), 276–282.

Doersch, C. (2016). Tutorial on Variational Autoencoders. *arXiv:1606.05908 [stat.ML]*, 1–23.

Du, M., Li, F., Zheng, G., and Srikumar, V. (2017). DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning. *Proc. of the ACM Conference on Computer and Communications Security*, 1285–1298.

Duda, R. O., Hart, P. E., and Stork, D. G. (Ed.). (2001). *Pattern Classification*. Wiley-Interscience.

Eid, A., Clerc, G., Mansouri, B., and Roux, S. (2021). A Novel Deep Clustering Method and Indicator for Time Series Soft Partitioning. *Energies*, *14*(5530), 1–19.

Elattar, H. M., Elminir, H. K., and Riad, A. M. (2016). Prognostics: a literature review. *Complex & Intelligent Systems*, *2*(2), 125–154.

Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., and Vincent, P. (2009). The Difficulty of Training Deep Architectures and the Effect of Unsupervised Pre-Training. *Proc. of the 12th International Conference on Artificial Intelligence and Statistics*, 153–160.

Farzad, A., and Gulliver, A. (2020). Unsupervised log message anomaly detection. *ICT Express*, *6*, 229–237.

Fefferman, C., Mitter, S., and Narayanan, H. (2016). Testing the Manifold Hypothesis. *Journal of the American Mathematical Society*, *29*(4), 983–1049.

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. (2020). Can Cross Entropy Loss Be Robust to Label Noise? *Proc. of the 29th International Joint Conference on Artificial Intelligence*, 2206–2212.

Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*(103678), 1–15.

Forest, F., Lebbah, M., Azzag, H., and Lacaille, J. (2019). Deep Embedded SOM: Joint Representation Learning and Self-Organization. *Proc. of the 27th European Symposium on Artificial Neural Networks*, 1–6.

Gelman, A. (2021). Reflections on Breiman's Two Cultures of Statistical Modeling. *Observational Studies*, *7*(1), 95–98.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, *10*(524), 1–15.

Gosset, W. S. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25.

Hancock, J. T., and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(28), 1–41.

Hernán, M. A., and Robins, J. M. (2020). *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Hu, X., Eklund, N., and Goebel, K. (2007). A Data Fusion Approach for Aircraft Engine Fault Diagnostics. *Proc. of ASME Turbo Expo*, *1*(GT2007-27941), 767–775.

Huang, B., Di, Y., Jin, C., and Lee, J. (2017). Review of Data-driven Prognostics and Health Management Techniques: Lessions Learned from PHM Data Challenge Competitions. *Proc. of the Conference of the Machine Failure Prevention Technology Society*, 1–17.

Huh, D. (2011). Synthetic Embedding-based Data Generation Methods for Student Performance. *arXiv:2101.00728 [cs.LG]*, 1–19.

Im, D. J., Ahn, S., Memisevic, R., and Bengio, Y. (2017). Denoising criterion for variational auto-encoding framework. *Proc. of the 31st AAAI Conference on Artificial Intelligence*, 2059–2065.

ISO. (2003). *Condition monitoring and diagnostics of machine systems: Data processing, communication and presentation* (Tech. Rep. No. 13374-1:2003). International Organization for Standardization.

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schölkopf, B. (2013). Quantifying causal influences. *The Annals of Statistics*, *41*(5), 2324–2358.

Kadra, A., Lindauer, M., Hutter, F., and Grabocka, J. (2021). Well-tuned Simple Nets Excel on Tabular Datasets. *Proc. of the 35th Conference on Neural Information Processing Systems*, 1–14.

Kanazawa, A., Sharma, A., and Jacobs, D. (2014). Locally Scale-Invariant Convolutional Neural Networks. *Proc. of the Twenty-eighth Conference on Neural Information Processing Systems: Deep Learning and Representation Learning Workshop*, 1–11.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised Contrastive Learning. *Proc. of the 34th Conference on Neural Information Processing Systems*, 1–23.

Kim, S., Choi, K., Choi, H.-S., Lee, B., and Yoon, S. (2022). Towards a Rigorous Evaluation of Time-series Anomaly Detection. *Proc. of the 36th AAAI Conference on Artificial Intelligence*, 7194–7201.

Kingma, D. P., and Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends (R) in Machine Learning*, 1–89.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, *4*, 3581–3589.

Lai, G., Chang, W.-C., Yang, Y., and Liu, H. (2018). Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. *Proc. of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1–11.

Lejeune, M. (2010). *Statistique, La théorie et ses applications*. Springer Verlag France.

Martinez, M., and Stiefelhagen, R. (2018). Taming the Cross Entropy Loss. *Proc. of the German Conference on Pattern Recognition*, 628–637.

Michau, G., and Fink, O. (2019). Unsupervised Fault Detection in Varying Operating Conditions. *Proc. of the IEEE International Conference on Prognostics and Health Management*, 1–11.

Molnar, C. (2019). *Interpretable Machine Learning*. Leanpub.

Patrini, G., Rozza, A., Menon, A., Nock, R., and Qu, L. (2017). Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Qiu, H., Eklund, N., Hu, X., Yan, W., and Iyer, N. (2008). Anomaly Detection using Data Clustering and Neural

Networks. *Proc. of the International Joint Conference on Neural Networks*, 3627–3633.

Rodriguez Garcia, G., Michau, G., Ducoffe, M., Sen Gupta, J., and Fink, O. (2021). Temporal signals to images: Monitoring the condition of industrial assets with deep learning image processing algorithms. *Proc. of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 1–13.

Runge, J., Bathiany, S., Bollt, E. *et al.* (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(2553), 1–13.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(eaau4996), 1–15.

Sammouri, W., Côme, E., Oukhellou, L., Aknin, P., and Fonlladosa, C.-E. (2014). Pattern recognition approach for the prediction of infrequent target events in floating train data sequences within a predictive maintenance framework. *Proc. of the IEEE 17th International Conference on Intelligent Transportation Systems*, 918–923.

Sârbu, S., and Malagò, L. (2019). Variational autoencoders trained with q-deformed lower bounds. *Proc. of the International Conference on Learning Representations*, 1–7.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards Causal Representation Learning. *Proc. of the IEEE*, *109*(5), 612–634.

Sejnowski, T. J. (2018). *The Deep Learning Revolution*. The MIT Press.

Shahid, N., and Ghosh, A. (2019). TrajecNets: Online Failure Evolution Analysis in 2D Space. *International Journal of Prognostics and Health Management*, *29*, 1–17.

Sicks, R., Korn, R., and Schwaar, S. (2020). A lower bound for the ELBO of the Bernoulli Variational Autoencoder. *arXiv:2003.11830 [cs.LG]*, 1–20.

Simard, P. Y., Steinkraus, D., and Platt, J. C. (2003). Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. *Proc. of the Seventh International Conference on Document Analysis and Recognition*, 958–962.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press.

Strickland, E. (2022). Are You Still Using Real Data to Train Your AI? *IEEE Spectrum*.

Szeliski, R. (2022). *Computer Vision: Algorithms and Applications*. Springer.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. *Proc. of the 25th International Conference on Machine Learning*, 2059–2065.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, *11*, 3371–3408.

Wu, R., and Keogh, E. (2021). Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. *IEEE Transactions on Knowledge and Data Engineering*, 1–9.

Xie, Y. J., Tsui, K. L., Xie, M., and Goh, T. N. (2010). Monitoring Time-between-Events for Health Management. *Proc. of the IEEE Prognostics and System Health Management Conference*, *MU3117*, 1–8.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2020). CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. *arXiv:2004.08697 [cs.LG]*, 1–21.

Yong, B. X., Pearce, T., and Brintrup, A. (2020). Bayesian Autoencoders: Analysing and Fixing the Bernoulli likelihood for Out-of-Distribution Detection. *Proc. of the 37th International Conference on Machine Learning*, 1–9.

Zaman, N., Apostolou, E., Li, Y., and Oister, K. (2022). Explainable AI for RAMS. *Proc. of the Annual Reliability and Maintainability Symposium*, 1–7.

Zečević, M., Dhami, D. S., Veličković, P., and Kersting, K. (2021). Relating Graph Neural Networks to Structural Causal Models. *arXiv:2109.04173 [cs.LG]*, 1–29.

BIOGRAPHIES

Alexandre Trilla graduated from La Salle University of Barcelona with a M.Sc. in Electrical Engineering in 2008, and a M.Sc. in IT Management in 2010. He has an academic research background in spoken language processing, and an industrial research background in PHM. He has authored several publications in scientific conferences and journals (International Journal of Prognostics and Health Management, IEEE Transactions on Audio, Speech, and Language Processing, Chemical Engineering Transactions, and the Journal of Rail and Rapid Transit). At present, he is a Senior Data Scientist and R&D Program Manager at Alstom, working on the deployment of PHM to the railway environment. He leads the development of predictive maintenance based on Machine Learning, and he is especially interested in building solutions using artificial neural networks and Deep Learning.

Nenad Mijatovic is a Data Science Leader in Alstom. He has over 20 years of algorithm development experience in a variety of areas, such as statistics, numerical optimization, machine learning, AI, and causality. Before joining Alstom, Dr. Mijatovic has held several R&D and leadership positions in the industry, from startups to blue-chip companies. His interests are applying machine learning and AI methods for industrial applications. In his current position, Dr. Mijatovic leads Alstom's data science teams responsible for delivering industrial-grade ML and AI algorithms for maintenance, operations, energy, and city flow solutions.

Xavier Vilasis-Cardona is full professor at La Salle, Universitat Ramon Llull, Barcelona. He holds a degree in physics ('89) and a PhD in physics ('93) by Universitat de Barcelona. He is member of the IEEE, of the IEEE CNNAC technical committee and of the LHCb collaboration. He is currently leading the Data Science for the Digital Society (DS4DS) research group.