

Neural Approaches to Prognostics and Health Management of Rolling Stock

Alexandre Trilla Castelló

<http://hdl.handle.net/10803/691418>

Data de defensa: 13-03-2024

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. The access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

DOCTORAL THESIS

Title Neural Approaches to Prognostics and Health Management of Rolling Stock

Presented by Alexandre Trilla Castelló

Centre La Salle Digital Engineering School

Department Engineering Department

Directed by Dr. Xavier Vilasís Cardona

Per a vosaltres, Ingrid i Gemma

ABSTRACT

Railway transportation is a mobility solution that must be both reliable and safe. To this end, the technical field of predictive maintenance focuses on applying data science to maximize the availability of rolling stock assets. This leads to modeling their degradation and minimizing their downtime by preventing service-affecting failures. To this end, Artificial Intelligence (AI) and Machine Learning have proven to be effective techniques for extracting latent patterns from the available data.

This dissertation puts the emphasis on Deep Learning, which is the state of the art in neural network research as the leading paradigm in AI and Machine Learning. Additionally, the scope of the work is framed in the multinational industrial context of Alstom, operating worldwide in rail markets, and active in the fields of passenger transportation, signaling and locomotives. The thesis is intended to be an expert reference work at Alstom in the area of predictive maintenance for rolling stock, especially through the use of neural networks for developing advanced maintenance solutions that are reliable and cost-effective. To this end, different environments have been considered, including mixed data types, i.e., continuous and discrete variables, and different predictive objectives such as diagnosis and prognosis. As a result of this research, three journal articles have been published (in addition to some conference papers).

RESUM

El transport ferroviari és una solució de mobilitat que ha de ser alhora fiable i segura. Amb aquesta finalitat, el manteniment predictiu se centra en aplicar la ciència de dades per a maximitzar la disponibilitat dels actius de material rodant. Això porta a modelar la seva degradació per a minimitzar el temps d'inactivitat mitjançant la prevenció de fallades que afecten el servei. En aquesta línia de valor afegit, la Intel·ligència Artificial (IA) i l'Aprenentatge Automàtic (AA) han demostrat ser tècniques efectives per a extreure patrons latents de comportament a partir de les dades disponibles.

Aquesta tesi posa l'èmfasi en l'Aprenentatge Profund, que és l'estat de l'art en la recerca de xarxes neuronals com a paradigma líder en IA i AA. A més, l'abast del treball s'emmarca en el context industrial multinacional d'Alstom, que opera en els principals mercats ferroviaris a nivell mundial, amb presència en els camps del transport de passatgers, senyalització i locomotores. La tesi pretén ser un treball de referència a Alstom en l'àmbit del manteniment predictiu del material rodant, especialment mitjançant l'ús de xarxes neuronals profundes per al desenvolupament de solucions de manteniment avançades que siguin fiables i efectives. Per a tal finalitat, s'han considerat diferents entorns amb tipus de dades mixtes, és a dir, amb variables contínues i discretes, i diferents objectius predictius com la diagnosi i la prognosi. Com a resultat d'aquesta investigació, s'han publicat tres articles en revistes indexades (a més d'algunes ponències en congressos).

RESUMEN

El transporte ferroviario es una solución de movilidad que debe ser a la vez fiable y segura. A tal fin, el mantenimiento predictivo se centra en aplicar la ciencia de datos para maximizar la disponibilidad de los activos de material rodante. Esto lleva a moldear su degradación para minimizar el tiempo de inactividad mediante la prevención de fallos que afectan al servicio. En esta línea de valor añadido, la Inteligencia Artificial (IA) y el Aprendizaje Automático (AA) han demostrado ser técnicas efectivas para extraer patrones latentes de comportamiento a partir de los datos disponibles.

Esta tesis pone el énfasis en el Aprendizaje Profundo, que es el estado del arte en la investigación de redes neuronales como paradigma líder en IA y AA. Además, el alcance del trabajo se enmarca en el contexto industrial multinacional de Alstom, que opera en los principales mercados ferroviarios a nivel mundial, con presencia en los campos del transporte de pasajeros, señalización y locomotoras. La tesis pretende ser un trabajo de referencia en Alstom en el ámbito del mantenimiento predictivo del material rodante, especialmente mediante el uso de redes neuronales profundas para el desarrollo de soluciones de mantenimiento avanzadas que sean fiables y efectivas. Para tal fin, se han considerado diferentes entornos con tipos de datos mixtos, es decir, con variables continuas y discretas, y diferentes objetivos predictivos como la diagnosis y la prognosis. Como resultado de esta investigación, se han publicado tres artículos en revistas indexadas (además de algunas ponencias en congresos).

ACKNOWLEDGEMENTS

I am deeply indebted to my supervisor, Prof. Xavier Vilasis, for his advice and sustained support throughout the development of my thesis. I will put it in clear counterfactual language: had he not encouraged me to enroll in the research program, this dissertation would not exist.

I would also like to have some warm words for my former adviser, Prof. Francesc Alías, who helped me get started in doing research and gave me the fundamental tools to succeed.

I would like to thank Prof. Olga Fink, Prof. Piero Baraldi, and Prof. Jordi Vitrià for chairing my PhD defense. Their feedback was invaluable to improve my dissertation.

I am utterly grateful to my managers and directors at Alstom, who have always supported my applied research endeavors, especially with respect to this dissertation: Vicente Fuerte, Sergi Bermejo, Guillermo Sospedra, Juan-Carlos Villalba, and Nenad Mijatovic. I would also like to thank my colleague co-authors, for the effort they put in improving the papers that we published.

I am also very grateful to the Government of Catalonia (Generalitat de Catalunya) for the financial support for a part of my PhD research project. This grant has enabled me to discover some of the greatest institutions and minds in science and learn from their knowledge and experience. Many thanks to all of the excellent professors and lecturers from La Salle in Barcelona, from Polimi in Milano, from Harvard in Boston, and from Northwestern in Chicago.

I would also like to have some kind words for my (then) comrades at La Salle, with whom I had my first steps into research and teaching... 16 years

ago! Thanks for the good time we had together Dr. Lluís Formiga, Dr. Santi Planet, and Prof. Xavier Sevillano.

Finally, last but not least, my family, Ingrid and Gemma, without whom some of the greatest events in my life would not have occurred.

CONTENTS

Contents	13
List of Tables	17
List of Figures	19
I Framework	1
1 Introduction	3
1.1 Motivation	5
1.2 Contribution and Hypothesis	6
1.3 Structure of the Dissertation	8
1.4 Published Work	9
1.4.1 Enhancing Railway Pantograph Carbon Strip Prognostics with Data Blending through a Time-Delay Neural Network Ensemble (2020)	9
1.4.2 Pushing Distributed Vibration Analysis to the Edge with a Low-Resolution Companding Autoencoder: Industrial IoT for PHM (2020)	10
1.4.3 Integrated Multiple-Defect Detection and Evaluation of Rail Wheel Tread Images using Convolutional Neural Networks (2021)	10

1.4.4	Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting (2022)	11
1.4.5	Unsupervised Probabilistic Anomaly Detection over Nominal Subsystem Events on a Hierarchical Variational Autoencoder (2023)	11
1.5	Acknowledgement	12
2	Railway-focused PHM	13
2.1	Main Topics in the Railway Research Community	14
2.1.1	Reliability	16
2.1.2	Management	16
2.1.3	Automation	18
2.1.4	Big Data	18
2.1.5	Cybersecurity	19
2.1.6	Sustainability	20
2.1.7	Wheel-Rail Interface	21
2.1.8	Brake	24
2.1.9	Bearings	24
2.1.10	Pantograph–Catenary Interface	26
2.2	Technologies Applied to Product Development	27
2.2.1	Smart Sensors	28
2.2.2	Machine Vision Inspection	29
2.2.3	Technical Language Processing	30
2.2.4	System Log Analytics	31
3	Deep Learning-based PHM	33
3.1	Main Topics in the Deep Learning Research Community	34
3.1.1	Breakthroughs in Neural Networks	34
3.1.2	Deep System Health Management	43
3.1.3	Challenges and Opportunities	44
3.2	Learning Techniques for Product Research	45
4	State Of The Art	47
4.1	Literature Review	48
4.1.1	Railway Engineering	48
4.1.2	Data Science	49
4.2	Railway Data Operations	50
4.3	Applied Research Questions	51

II Contributions	55
5 Research Publications	57
5.1 Conference Paper 1 (2020)	59
5.2 Conference Paper 2 (2020)	71
5.3 Journal Article 1 (2021)	83
5.4 Journal Article 2 (2022)	105
5.5 Journal Article 3 (2023)	125
6 Discussion and Conclusions	143
6.1 Interpretability and Explainability	144
6.2 Decision Making	145
6.3 Industrialization	145
6.4 Causal Inference	146
6.5 Updated Challenges and Opportunities	148
6.5.1 Railway Fleet Planning	150
6.5.2 Technical Language Processing in Retrospect	150
6.6 On a Final Note...	151
Bibliography	153
Index	177

LIST OF TABLES

- 4.1 Potential impacts of the AI technology in the railway maintenance industry and business in general. 50
- 6.1 Aggregated challenges and opportunities along with the Conference Papers (CP) and Journal Articles (JA) that addressed them, also showing the publication years in brackets. 149

LIST OF FIGURES

- 3.1 Multilayer Perceptron network where $O = g(I \cdot W_{IH}) \cdot W_{HO}$. The non-linear function g is inherent in the hidden layer. In this diagram, the bias terms “+1” are made explicit. 35
- 3.2 Autoencoder architecture, where D is the data dimensionality and H is the size of the hidden layer, which defines the representational capacity of the network. The compressive encoding function of the model is ensured as long as $H < D$. 36
- 3.3 Convolutional layer for an input vector $I = (i_1, i_2, \dots)$, a 1D second-order filter $W = (w_1, w_2, w_3)$, and an output vector $O = (o_1, o_2, \dots)$, which clearly shows that $O = I * W$. Edge thickness indicates parameter reuse. 38
- 4.1 Map of railway assets that appeal to business (in circles) and their related predictive maintenance technologies (in squares). 51

PART I

FRAMEWORK

INTRODUCTION

We're in the business of demonstrating a learning capability, showing that something can learn to do something nontrivial and that it can learn it whether it's a problem that's amenable to analytical treatment or not.
– Bernard Widrow (1994)

THE Fourth Industrial Revolution and digital technologies are major drivers of innovation. These incentivize businesses to maximize productivity, which in turn spur economic growth (Jahan, S. and Mahmud, A. S., 2015). However, such growth can eventually lead to an exhaustion of the available resources, which are already scarce by definition, and ultimately increase social inequality, thus affecting the human talent that runs the companies. As a consequence of this risk, increasingly more sustainable and efficient alternatives lead the agenda of corporations and academia.

INNOVATION

In light of these challenges, Artificial Intelligence (AI) is regarded as a field in science and engineering that can help to optimize the capacity of business, but it is imperative for leaders to separate AI reality from hype. Progress in AI is taking place in a technological context marked by the datafication of the world which affects all sectors of our society and economy. Nonetheless, if there is one strategic area that is particularly well suited to integration into AI it is the transport and mobility sector, which is one of Europe's longstanding strengths (Villani, C., 2018). In fact, if we

ARTIFICIAL INTELLIGENCE

TRANSPORT

focus our attention on the mass transit environments, there is a push by Europe's major railways and manufacturers to adopt new technology through the use of AI and the Industrial Internet of Things (Scordamaglia, D., 2019). Thus, the digitalization is going to play an important role in helping railways to carry more passengers and freight without having to invest huge sums of money in the time-consuming process of laying additional tracks and expanding stations (Briginshaw, D., 2020a).

PREDICTIVE MAINTENANCE focuses on the application of AI to the predictive maintenance of rolling stock, which is meant to maximize the availability of these transport assets. The impact of such digitalization will be a game-changer on all maintenance activities in the railway transport sector (UITP, 2020). By employing advanced software and engineering techniques, predictive technology can substantially enhance railway safety by enabling a shift from time-based to needs-based maintenance (Man, T., 2018). This should improve operational availability and efficiency while reducing maintenance costs. Analyzing data from trains in service provides insights into their degradation, which helps maintainers make better informed decisions to take action on their fleet given the limited resources at the depot. Moreover, the extracted information enables creating several new business cases. For example, value is added when the whole life of the components is used. This does not occur when the traditional over-dimensioned scheduled maintenance process is followed. In that case, there are many frequent operations that are probably unnecessary because the replacements are usually planned at a fraction of their expected life driven by risk-averse criteria, regulation, and conservative supplier policies. Additionally, value is also added when unexpected issues occur and they are detected ahead of time (and fixed early) at the point of incipient failure, so that the catastrophic expense of a potential service-affecting failure is avoided. All these refined features also lead to an increase of safety in the railway transport service (Seisenberger, M., ter Beek, M. H., Fan, X., Ferrari, A., Haxthausen, A. E., James, P., Lawrence, A., Luttkik, B., van de Pol, J., and Wimmer, S., 2022), which is likewise subject to business profitability.

RAILWAY

BUSINESS CASE

SAFETY

DIAGNOSIS
PROGNOSIS

In technical terms, the goal of predictive maintenance is framed under the scope of Prognostics and Health Management (PHM), which is the proper scientific field that formally studies these topics. Broadly speaking, PHM is a data-driven approach that provides insights into the actual health condition of a degrading asset (i.e., the diagnosis) and predicts its future evolution (i.e., the prognosis) as an estimation of the remaining useful life. These concepts were originally introduced by Hippocrates (c. 460 – c. 370

BC) in the context of medicine (Stefanakis, G., Nyktari, V., Papaioannou, A., and Askitopoulou, H., 2020). At present, the bleeding edge of research in PHM is fueled by the recent successes of Deep Learning (DL) on many scenarios (Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020). DL approaches are able to process raw data directly and learn representations on many levels. This is especially appealing nowadays in the maintenance environment because tackling the problems following the conventional engineering approach, which is advised by a committee of subject matter experts, is not feasible anymore (Smith, K., 2023). The fast-paced increasing sophistication of the assets and their changing operational regimes and environments, in addition to external factors such as the energy crisis (Smith, K., 2022a), the war in Ukraine (Clinnick, R., 2022a), and the Covid-19 pandemic (Burroughs, D., 2022d), put intense pressure on maintenance and engineering teams and threaten the delivery of rail projects.

The overall purpose of this research is to capitalize the enhanced data-driven power provided by PHM through the DL technology, and use it to add value to Alstom's rolling stock maintenance business through the pursuit of sustainable and efficient solutions. To this end, additional challenges need to be tackled, such as the discovery of anomalous behaviors in regular service data without a record of previous failures, while dealing with different sources of data (e.g., rich parametric signals from sensors compared to sparse nominal variables from subsystem events). This Chapter is organized as follows: Section 1.1 describes the motivation of the research, Section 1.2 details the contributions, Section 1.3 outlines the structure of the dissertation, and Section 1.4 summarizes the published work.

1.1 Motivation

Richard Hamming, the renown American mathematician for his contributions is computer engineering and telecommunications, argued that if you are to do important work then you must work on the right problem, at the right time, and in the right way (Hamming, R. W., 1986). Valuable work is not simply new/original work that no-one cares about, it must be relevant and impactful, especially in a vibrant field like DL research (Wagstaff, K. L., 2012).

The industrial research conducted in this dissertation is regarded to be important for the following reasons:

Right problem The problems tackled here align with Alstom's maintenance business. The business case for PHM yields a positive return on investment when the predictive technology is applied to components

that rarely fail but which are very costly to repair, which is the case for bogie components such as axle bearings, wheels, traction motor, etc.

Right time Alstom has recently been through a process of deep transformation by merging with Bombardier Transportation to enhance its global presence and become the leading innovator in the market (Poupart-Lafarge, H., and Smith, K., 2021). This research aligns perfectly with this vision.

Right way The fundamental role of a research scientist at Alstom is to capitalize the state of the art in PHM and apply it to specific product/business problems, rather than delving into the techniques and competing against the best performing approach in the literature (provided that it is the same problem, data, and performance indicators). The developed solutions must be good enough to adapt fast to the pace of change and continuously add value to the customers in an agile manner. The Right First Time management principle is a way-of-working recommendation that has traditionally been observed at Alstom to make sure these points are effectively regarded (Leuenberger, H., Puchkov, M., and Schneider, B., 2013).

ALSTOM Having Alstom as the committed industrial partner ensures that the research is to the point, fit for purpose, and of interest to a large community of railway engineers. Note that a rolling-stock manufacturing company like Alstom does not take for granted that data science adds value to its core business. That is why it is important to focus on the application of the research, which in this case its main contribution is in the use of DL for tackling the variety of data through the developed solutions on 5 existing data products.

1.2 Contribution and Hypothesis

In the railway sector, neural network techniques have been used to develop condition monitoring solutions since the mid nineties (Fararoy, S., and Allan, J., 1995). The neural technology that was available at the time is now regarded as the 2nd generation of networks, also known as Multilayer Perceptrons. They are able to learn and approximate any continuous non-linear function using gradient descent optimization strategies, but their capacity becomes quickly limited as more representational power is pursued. Nowadays, a 3rd generation of networks known as Deep Learning has superseded the Multilayer Perceptrons, and more intricate objective functions can now

be learnt through the depth of the models and the parallelization of their computations.

This dissertation is intended to be an expert reference work at Alstom in the field of predictive maintenance for passenger rolling stock, especially through the use of Deep Learning as the state of the art technology in the study of neural networks. The following list describes the goals that are expected to be attained through the different solutions developed in this research:

Cost-effectiveness Prove that DL is the most suitable approach to tackle PHM problems at Alstom, i.e., a solution that is good enough, developed in a short time, and thus able to increase the productivity.

Proven technology Prove that the DL technology of use stands the test of time, i.e., the applied research shows more than 5 years of sustained progress, showing a stable trend.

Industrialisability Prove that the lead time between a research prototype and an industrial-proof solution on any platform is minimized with DL.

Flexibility Prove that the DL technique can be applied to the whole value chain of PHM for solving different problems in terms of components and variables, and also to improve the performance of the related products.

Robustness Prove that DL succeeds in solving real-world railway problems, characterized by a shortage of critical failure data, which is very different from standard research datasets.

Being aware that there are multiple ways/approaches to accomplish a given goal, and that a company like Alstom is strongly rooted in traditional engineering ways of working, which sometimes raise qualms about recent research progress, this dissertation should build a convincing case for applying DL to the many PHM challenges in the maintenance of railways. This mindset often leads to preferring the development of a single solution as a specific proof of concept rather than pursuing vast comparisons with other previous work, which may also be unfair due to lacking implementation details. Such inherent limited reproducibility may also lead to cherry-pick results in order to make a questionable point, intentionally or inadvertently (Komiyama, J., and Maehara, T., 2018). Moreover, the huge expressiveness of DL yields models that can be trained to learn virtually anything,

including totally meaningless associations with shuffled labels (Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O., 2017), which could lead to wrong conclusions. Additionally, the bitter lesson that the general methods that leverage computation are ultimately the most effective ones (instead of other factors such as the refinement of the algorithms) leads to conclude that the approaches that are blindly focused on performance comparisons are likely to be rapidly superseded by more powerful computers (Sutton, R. S., 2019). Finally, the fitness for industrialization often filters the range of approaches that may even be worth comparing, in addition to the benefit of productionizing well time-tested interpretable techniques, which is obviously unattainable when the bleeding edge of the state of the art is targeted.

HYPOTHESIS

Taking into account all these considerations, the research hypothesis that is here tackled can be expressed as follows:

Deep Learning displays the characteristics that make it a suitable technology for developing dependable industrial-grade solutions for effectively maintaining rolling stock with confidence

This dissertation deals with the most recent research in this line of work at Alstom, which covers the past 3 years, i.e., from 2020 until now. However, this effort could be framed in a broader scope that stretches back to the last 10 years (Trilla, A., and Gratacòs, P., 2013, 2016; Trilla, A., Gratacòs, P., Guinart, D., Alessi, A. and Lamoureux, B., 2016; Trilla, A., and Cabré, X., 2018; Trilla, A., Dersin, P., and Cabré, X., 2018; Trilla, A., Janjua, F., and Bermejo, S., 2019).

1.3 Structure of the Dissertation

This dissertation is divided into two main parts. Part I introduces the framework in Alstom's industrial context and develops the state of the art, and Part II provides the published scientific contributions, discusses their impact, and draws the conclusions.

In more details, Part I is organized as follows: Chapter 1 provides the motivation for this research, the corporate business context where it is developed, and its hypothesis. Chapter 2 develops the state of the art from the viewpoint of the railway engineer, including general topics such as reliability and management, among others, and also focuses on their application to product development. Chapter 3, instead, adopts the perspective of the data scientist, and focuses on the technical topics around Deep Learning and the application of specific approaches to PHM.

Similarly, Part II is organized as follows: Chapter 5 provides a summary of the five published peer-reviewed scientific contributions in the recent research period, which comprises two conference papers and three journal articles. Finally, Chapter 6 concludes the work, discusses the limitations of the research, asks challenging questions about the interpretability of the results, their value in helping make decisions, and describes interesting lines of future improvement, mostly observing the inference of causality.

1.4 Published Work

This dissertation presents a Ph.D. Thesis as a compendium of publications where the candidate has had a *leading role* in all of them, comprising their conceptualization, experimentation, writing, review and editing. All the studies and investigations comprised in this research started as specific problem to be solved in a project, and the candidate developed bespoke solutions based on Deep Learning that align with the state of the art, resulting in publications that met the quality standards of scientific media. Therefore, his contributions constitute a solid base on which to build products that enhance the predictive maintenance of rolling stock. This Chapter summarizes the two conference papers (2020) and the three journal articles (2021–2023) that have been published in this doctoral period.

1.4.1 Enhancing Railway Pantograph Carbon Strip Prognostics with Data Blending through a Time-Delay Neural Network Ensemble (2020)

This contribution develops a straightforward solution to industrialize the prognostics of pantograph degradation based on the thickness of the carbon strips. The predictive method is based on a robust online non-linear multivariate regression approach, and considers factors that may have an impact on the degradation on the carbon strip, such as the seasonal condition of the overhead contact wire. Its implementation is based on a neural ensemble using a Multilayer Perceptron. The learning approach aims to integrate all the sources of potential utility along with the carbon strip data, which is averaged in time with a set of spreading filters to increase the overall robustness to uneven sampling. Finally, the uncertainty of this technique is determined with a sliding window approach, and the resulting accuracy is ensured to be within the specifications for adding value to the maintenance of the rolling stock, i.e., a small confidence interval for a given horizon than enables the team to schedule the resources at the depot ahead of time.

1.4.2 Pushing Distributed Vibration Analysis to the Edge with a Low-Resolution Companding Autoencoder: Industrial IoT for PHM (2020)

This contribution explores a vibration data compression strategy for diagnosis purposes. This work is motivated by the low-bandwidth transmission capacity of the radio interfaces that wireless sensor networks typically equip, and the low-power features of their battery-operated (and/or energy-harvested) electronics. The proposed approach first compresses the raw signal waveform using an optimally regularized Autoencoder with an undercomplete representation, and then it reduces the resolution of the compressed data by quantizing all the resulting real values into single-byte unsigned integers. The evaluation of this strategy on a dataset of railway axle bearings has concluded that with compression rates up to 10 the vibration signals are practically unaffected by this procedure, and once the signals are reconstructed, many diagnosis goals like anomaly detection, fault location, and severity appraisal can be performed. Moreover, the obfuscated embedding of the compression may be seen as a means to encrypt the data for cybersecurity purposes, especially if more depth is considered in the Autoencoder.

1.4.3 Integrated Multiple-Defect Detection and Evaluation of Rail Wheel Tread Images using Convolutional Neural Networks (2021)

This contribution presents an automatic Deep Learning method to jointly detect and diagnose wheel tread defects based on smartphone pictures taken by the maintenance team on the shop floor. This approach is based on a framework of Convolutional Neural Networks, which is applied to the different tasks of the diagnosis process including the location of the defect area within the image, the prediction of the defect size, and the identification of the defect type. With this information determined, the maintenance-criteria rules can ultimately be applied to obtain the actionable results. This work concludes that the presented method can reliably automate the condition diagnosis of half of the current workload and thus reduce the lead time to take maintenance action, significantly reducing engineering hours for verification and validation.

1.4.4 Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting (2022)

This contribution explores how the Causal Inference paradigm may be applied to troubleshoot the root causes of failures through language processing and Deep Learning. Following the leading state-of-the-art strategy to represent latent features from text (Feder, A., Keith, K. A., *et al.*, 2022), a novel approach to extract linguistic knowledge has been devised through the joint embedding of two contextualized Bag-Of-Words models, which defines both a probabilistic framework and a distributed representation of the underlying causal semantics. This method has been applied to the maintenance of rolling stock bogies using Return On Experience data, and the results indicate that the inference of causality has been partially attained with the currently available technical documentation (consensus with failure analysis over 70%). Additionally, the proposed approach may be used as a strategy to detect lexical imprecision, make writing recommendations in the form of standard reporting guidelines, and ultimately help produce clearer diagnosis materials. As a result, the safety of the railway service may be increased by flagging ambiguous expressions and words that could cause communication errors (Nakamura, R., 2019).

1.4.5 Unsupervised Probabilistic Anomaly Detection over Nominal Subsystem Events on a Hierarchical Variational Autoencoder (2023)

This contribution develops a versatile approach to discover anomalies in massive operational data for nominal (i.e., non-parametric) subsystem event signals using unsupervised Deep Learning techniques. Firstly, the proposed method builds a neural convolutional framework to extract both intrasubsystem and intersubsystem patterns. Secondly, it generalizes the learned embedded regularity of a Variational Autoencoder manifold by merging latent space-overlapping deviations with non-overlapping synthetic irregularities. Finally, it creates a smooth diagnosis probabilistic function on the ensuing low-dimensional distributed representation using a Multilayer Perceptron. 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. Additionally, its results yield interesting opportunities for designing Intrusion Detection Systems in the context of cybersecurity.

1.5 Acknowledgement

Amb el suport del Pla de Doctorats Industrials de la Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya (Referència 2020 DI 54).

RAILWAY-FOCUSED PHM

Perhaps coming up with a theory of information and its processing is a bit like building a transcontinental railway. You can start in the east, trying to understand how agents can process anything, and head west. Or you can start in the west, with trying to understand what information is, and then head east. One hopes that these tracks will meet.
– John Barwise (1986)

THE purpose of maintenance is to keep assets performing their prescribed functions at the optimum cost (UITP, 2020). Due to competitive pressure and overcapacity in production facilities, some rolling stock manufacturers seek growth in new business models, e.g., offering train-as-a-service models (McKinsey, 2017). To attain this goal, sensor technology and data analytics have to change the maintenance paradigm: from time and usage-based maintenance to condition-based and predictive maintenance. The adoption of this paradigm has the following benefits (UITP, 2020):

BUSINESS MODEL

- Faster identification and timely qualification of asset deterioration
- Increased asset availability and optimized maintainability for the operators
- Improved asset reliability and safety, leading to more trust from passengers and better reputation for the operator

- Lower system life cycle costs

INSPECTION The costs of the condition-based models are mainly influenced by the frequency of inspections. But if there aren't any inspections performed, the costs are going to be higher due to the rising number of failures (Eisenberger, D., and Fink, O., 2017). Going from corrective, to planned, to condition-based, and to predictive maintenance, reliability increases and cost decreases. Similarly, going from reactive, to manual, to automated, and to data-driven procedures, the required effort increases as well as potential benefit (Thompson, I., 2022). However, research and discussion with asset designers, manufacturers, owners, operators and maintainers showed that railway-focused condition-based maintenance (CBM), as a general cohesive concept, is still in its infancy (UITP, 2020).

IMPACT Railway researchers are urged to make an impact. For any technological transition to be successful, it is essential that everyone is heading in the right direction (Smith, K., 2022c). Data on recent patent applications made around the world provides insights into the rail innovative technology leaders (Clark, M., 2022), and in this regard, the number of Chinese applicants has risen significantly in recent years and is likely to continue to do so. Additionally, the Shift2Rail program aims to deliver, through railway research and innovation, the capabilities to bring about the most sustainable, cost-efficient, high-performing, time-driven, digital and competitive customer-centered transport mode for Europe (Shift2Rail, 2020).

Research shows that maintenance performance is linked to many heterogeneous parameters, and that most of them are not yet taken into account in the current maintenance processes (Unife, 2017). This chapter describes some important aspects from the viewpoint of the railway engineer. First, Section 2.1 focuses on the hot topics around railway maintenance, along with the different techniques to tackle these challenges, including model-based, data-driven, and hybrid approaches (Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N., 2017). Then, Section 2.2 provides the outlines of the technologies that are used to develop solutions that address the former questions in different environments.

2.1 Main Topics in the Railway Research Community

VALUE CHAIN While it is difficult to determine exactly how fast and how far AI will go in terms of disrupting the business value chain, there is a consensus that the ability to automate processes, analyze data beyond human comprehension,

and personalize customer services will have profound and far-reaching impacts on how companies operate (Barrow, K., 2018b). What follows is a list of potential impacts of the AI technology in the railway maintenance industry (Burroughs, D., 2019c) and business in general (Glover, J., 2013):

- Increased capacity: railway companies are paid to provide movement, and load factors are all-important, i.e., the trains must be full and they must be used for a large proportion of each day and every day. Line capacity is a scarce resource, and spare capacity is a drain on company resources.
- Reduced life cycle costs: short turnarounds are key to utilization.
- Reduced errors from both humans and existing computer systems. Good performance is vital: performance of passenger trains is measured by punctuality and reliability, and the railway must be seen as safe.
- Improved efficiency and increased performance: faster journey times allow better use of staff and stock.
- High-level automation and autoadaptive systems: rail traffics are interdependent, i.e., the railway does not exist in a vacuum.
- Simplified supervision and fast problem resolution, which is driven by reduced complexity with interoperable interfaces.
- Improved flexibility, taking into account that change takes time, but that it is inevitable because continuous improvement updates are necessary.

Remote diagnosis and CBM are considered a major change in train maintenance. They present a significant opportunity to reduce maintenance costs, while also having a positive impact on reliability, availability and service quality (Verdun, C., Audier, P., and Turgis, F., 2020). CBM can reduce rolling stock manual diagnostics by at least 60% and could lead to an overall reduction of at least 10-15% in maintenance costs: equivalent to an annual saving of up to 4bn euro for train operators, 2bn euro for rolling stock manufacturers, and 4bn euro for third parties (McKinsey, 2017). Thus, CBM is a key business driver for digitalization in the rail sector (Barrow, K., 2018a).

REDUCTION

2.1.1 Reliability

Reliability is the key to customer satisfaction in public transport, and perceptions of reliability increase when passengers have early warning of unexpected delays thanks to the provision of real-time disruption information (Blome, C., Kargoll, B., and Wernz, J., 2022). Evidently, safety must always be a priority, coming before all others. But other modes of transport appear to be able to balance the need for safety while also implementing new technologies at a rate faster than rail (Burroughs, D., 2022b).

RELIABILITY The benefits of higher reliability and lower costs can only be realized when the organizational culture adapts to make best use of the technology (Kilian, K., Kilian, M., Mazur, V., and Phelan, J., 2016). Reliability Centered Maintenance (RCM) and Life Cycle Cost allow factoring a cost-benefit analysis into the technical considerations to yield the optimal final maintenance decision (Mascherona, R., Bellani, L., Compare, M., Trucco, R., Zio, E., 2020). However, the maintenance concept has become so mistakenly entwined with reliability that the two terms are often used as synonyms. On the one hand, reliability focuses on failure patterns: bathtub, wear out, fatigue, initial break-in period, infant mortality, etc. (O’Hanlon, T., 2019). On the other hand, maintenance focuses on the return on assets: chronic failure analysis, RCM analysis, root-cause analysis, defect elimination, integrated cost-schedule management, work-order feedback, etc. (Reeve, J., 2019).

COST-BENEFIT

FAILURE Anyhow, failures are an undeniable part of maintenance. Failures happen irrespective of the strategies implemented for their prevention, whether based on reliability or maintenance. However, with a better planning approach, these failures should be used to improve the state of the system. By employing the Pareto principle, which states that 80 percent of the issues comes from 20 percent of the causes, any organization can focus its attention on genuine and demanding issues related to maintenance and reliability, rather than overreacting on each failure (Khan, S., and Yairi, T., 2018).

2.1.2 Management

Asset Management

Many large rail projects are doomed to failure and, if they are realized, they are likely to be completed with unexpectedly high costs and rarely without defects (van der Bijl, R., Utsunomiya, K., and van Oort, N., 2020). Therefore, predictive maintenance remains the holy grail for effective railway management (Smith, K., 2022b). The emerging digital technologies and AI are expected to augment the decision making in asset and fleet management.

However, the AI technologies need to be adapted to the specific needs of any industrial domain, and facilitate the implementation and achievement of the overall business goals (Kumari, J., Karim, R., Thaduri, A., and Castano, M., 2022).

Traditional solutions rely on constant signaling thresholds defined using technical knowledge and physical models to assess the health state of a system. As the health of a system differs from one train to another and independently evolves in time, these thresholds do not always take into account maintenance load, maintenance infrastructure availability and the effect of aging during the lifetime of the system. To overcome this limitation, a hybrid system mixing system experts and machine learning tools is advised (Turgis, F., Audier, P., Nemoz, V., and Marion, R., 2022). For example, statistical fleet analysis could be combined with clustering algorithms to characterize the health state of a system and identify potential failure modes. This approach could help to predict how aging effect and operational constraints impact the wear of a system.

THRESHOLD

AGING

WEAR

Here, the opportunity is to automate decision-making and decision implementation with the support from subject matter experts (Apps, J., 2019). In this sense, an AI-based system could support inexperienced staff to make better and more informed decisions. It may also help reduce the stress by allowing staff members to concentrate on other priority tasks (Clinnick, R., 2022b). Eventually, maintenance development and fleet maintenance management will potentially merge into a single maintenance analytics and scheduling function (McKinsey, 2017). Creating and updating decision rules and implementing them in the maintenance processes will be at the heart of the new maintenance system.

DECISION-MAKING

RULES

Talent Management

The rail sector overall, including maintenance and operations, is responsible for more than 1 million direct and 1.2 million indirect jobs in the EU (Shift2Rail, 2020). Understanding exactly when a maintenance intervention is required can optimize the allocation of human resources, reducing the overall need for the most skilled technicians who are increasingly hard to come by (Smith, K., 2022b).

JOBS

HUMAN RESOURCES

People are key to making PHM a success (Burroughs, D., 2019a). However, as the nature of work changes and the useful life-span of skills drops, training departments are facing the challenge of producing more new staff with digital skills as well as more practical skills to meet the needs of railways (Burroughs, D., 2019d). In the future there are likely to be four categories of tasks comprising those operated by only humans, humans

and machines (both separated as well as integrated into AI), and only machines/AI (Burroughs, D., 2019d). Therefore, one of the biggest challenges is to provide change management support and training for people (Verdun, C., Audier, P., and Turgis, F., 2020).

TRAINING

2.1.3 Automation

Automatic Train Operation, the technology behind driverless and unattended trains, can change the way the stock is maintained. Succinctly, the vehicle receives the distances it is allowed to drive and its permissible speeds via radio signals and makes sure these are complied (Clinnick, R., 2021a). Autonomous driving along the route and within the depot relieves the drivers and increases the safety of the passengers and other road users (Clinnick, R., 2021c). Moreover, real-time passenger flow prediction and crowd management could also be used to improve dwell time performance at key nodes in the network (Le Glatin, N., and Clarke, P., 2021). All these additional data may be used to infer the expected degradation and thus improve the management of the maintenance operations. As an example of this change of paradigm, the phased construction of Barcelona's first fully-automated metro line (i.e., L9, using the Serie 9000 fleet, constructed by Alstom) has been accompanied by a complete rethink in the way the entire network is operated and maintained, and this has produced several business benefits (Briginshaw, D., 2017).

AUTONOMOUS DRIVING

PASSENGER FLOW

DWELL TIME

2.1.4 Big Data

Data alone is only cost. The real benefit comes when you can turn it into insights (Barrow, K., 2018a). Thus, applications and AI could be the answer to creating a better public transport within cities and reducing the dependence on cars. Mobility must be considered a technical as well as a social issue and focus on people's requirements. However, integrating more services, more data sources and more platforms makes the whole management of a mobility system increasingly complex (Clinnick, R., 2022b). Mobility-As-A-Service should promote: the use of a single app to provide access to mobility and custom payment, the facilitation of a diverse menu of transport options, a competitive alternative to the private use of cars, and digitalization as an aid to the effectiveness of the transport system (Barrow, K., 2019a). Affordability, ease of access, and straightforward journey planning are key to changing travel patterns, and these are bound to change the approach to maintain the stock and ensure that it is available to meet its dynamic demand.

INSIGHTS

MOBILITY

TRAVEL PATTERN

Finally, the way data is collected must be carefully considered. Monitoring assets in testing mode can provide results that are not accurate, since the information are coming from assets that are not in their operating conditions. Instead, monitoring the assets from a train that is in regular service can help addressing this issue (Derosa, S., Frøseth, G. T., Lau, A., and Rönquist, A., 2022). When it comes to managing large amounts of data, there are two kinds of approaches: those that start out building a scalable infrastructure, and those that are in business (Helland, P., 2020).

OPERATING CONDITION

2.1.5 Cybersecurity

With the adoption of information and communications technologies in railway maintenance, vulnerability to cyber threats has increased. It is essential that organizations move toward security analytics and automation to improve and prevent security breaches and to quickly identify and respond to security events (Kour, R., Aljumaili, M., Karim, R., and Tretten, P., 2019). In this sense, the rail sector shows low levels of maturity compared with other sectors such as aviation (Burroughs, D., 2023). The current concern is the existing installed base, because that's what is transporting passengers at present. The characteristics of railway networks that make them a potential target include a distributed architecture, long life-cycles for equipment, high safety integrity levels (redundancy), diversity of supply chain, and small-medium volume production (Barrow, K., 2018c).

SECURITY

Increasing digitalization opens the railway up to a broad range of cybersecurity threats, both known and unknown. The risks of not taking data protection seriously include loss of intellectual property, the theft of sensitive data, and damage to high value systems and infrastructure (Burroughs, D., 2020). The most likely threats against them are Denial-of-Service (DoS) attacks, which are designed to shut down a machine or network, making it inaccessible to its intended users. This is usually achieved by flooding the target with traffic, or sending it information that triggers a crash.

PROTECTION

DENIAL-OF-SERVICE

Cyberattacks are also increasing in railways with an impact on railway stakeholders, e.g., threat to the safety of employees, passengers, or the public in general; loss of sensitive railway information; reputational damage; monetary loss; erroneous decisions; loss of dependability, etc. (Kour, R., Karim, R., and Thaduri, A., 2020). Missing awareness is one of the biggest issues in the railway domain (Burroughs, D., 2019b).

While standard engineering approaches are effective in building new rail control system components, a broader and more creative consideration of attacks has benefits. In particular, the ability to cause mass disruption by targeting the fail-safes designed to ensure safety or auxiliary systems that are

not directly classified within the scope of the industrial control systems (Unwin, D., and Sanzogni, L., 2022), which are the many unseen but important cogs in the world that control the critical railway infrastructure (Villareal, J. F., 2019). On this point, the real problem with cybersecurity comes from telecommunications (Briginshaw, D., 2019), and there the most common cyberattack in the transportation and rail sector comes from malware (Kour, R., Aljumaili, M., Karim, R., and Tretten, P., 2019). In this sense, signaling systems are on the spotlight (Briginshaw, D., 2020b), especially when the goal is to standardize them for interoperability purposes between countries (Briginshaw, D., 2022b; Rodenbeck, A., and Clinnick, R., 2022).

MALWARE

Alstom's stake in rail cybersecurity company Cylus is an indication that rail has moved into the software age. Cyberattacks have increased by 173% since 2016 to the point where there is now, on average, a cyberattack against critical rail systems every 30 days. Cybersecurity will become the number one priority for top management alongside strategy (Briginshaw, D., 2022a), which includes: 1) security by design, separating security and safety functions; 2) reduced attack surface, with minimum physical and functional interfaces; and 3) defense through depth, building multilayered mechanisms for security and detection (Burroughs, D., 2021b).

CRYPTOGRAPHY

Finally, security issues have posed serious challenges for the widespread application of AI. Cryptography is the core technology to solve security problems, and how to adapt it to AI is a key issue. The state-of-the-art mainly focuses on secure multiparty computation, homomorphic encryption, secure outsourcing computation, and federated learning. In addition, verifiable technology has also become important to ensure the correctness and integrity of AI systems. However, some solutions are high consuming in computation or communication, which greatly impacts the usability and practicability. Thus, exploring lightweight cryptographic techniques for AI is a challenging research direction (He, D., 2023), and using AI to detect intrusions is a topic that is becoming more and more relevant in the rail industry as the IoT becomes more ingrained in systems and processes (Burroughs, D., 2019c).

2.1.6 Sustainability

OVERHAUL

In railways, 9-12% of total vehicle operating costs are spent on bogie maintenance and life-cycles can be extended significantly (i.e., overhaul intervals by 25-75%) with the aid of remote condition monitoring (Barrow, K., 2018a). Therefore, using the whole life of a component and maintaining the stock only when needed (in contrast to a time-based schedule) contributes

SUSTAINABLE

to a more sustainable maintenance service.

The EU estimates that pollution, CO₂ emissions, noise and congestion costs the EU 1bn euro annually (Briginshaw, D., 2021), and rail traction produces about 2.9m tonnes of CO₂ (Cooney, N., 2020). In this line of sustainability, there appears to be three clear choices for decarbonizing the railway: electrification, batteries and hydrogen (Clinnick, R., 2021b). However, policymakers don't see rail as a solution to the climate crisis and discussions are focusing too much on new technological solutions like electric cars (Smith, K., 2021). Rail is the most energy-efficient and environmentally-friendly form of powered transport. This is why European strategies for CO₂ reduction see a lot of potential in supporting further railway development, but strong leadership is needed (Briginshaw, D., 2020a)

DECARBONIZING

CO₂

Rail systems make a compelling case for being at the core of any “net-zero” future transport system: full electrification (with a fixed infrastructure), huge capacity potential and low energy-loss running dynamics (Ward, C., Goodall, R., Harrison, T., and Midgley, W., 2022). Rail is making strides to become more sustainable in its operations, but work remains to reduce the carbon footprint of constructing new lines and stations (Burroughs, D., 2022c).

ELECTRIFICATION

Alternatively, rail vehicle light-weighting using fiber reinforced polymer composite materials is essential for the future of rail. This is recognized as a means of reducing carbon dioxide production through lower energy consumption, as well as reducing the impact on track degradation, thus delivering improved rail capacity and performance (Bruni, S., Mistry, P. J., Johnson, M. S., *et al.*, 2022).

Finally, battery traction is beginning to come into its own as a viable alternative to electrification and diesel (Burroughs, D., 2022a). Batteries are believed to be the best alternative way of powering rolling stock (Barr, A., and Smith, K., 2022). Also, they can solve AC-DC transfer issues by avoiding expensive infrastructure changes and simplifying the track layout (Hameed, R., 2021). Nevertheless, electric multiple units only do 50-80km on battery. Thus, hydrogen is the only technological solution at present that will get anywhere near providing autonomous traction. In this sense, Alstom led the charge to introduce hydrogen in Europe in 2016 (Clinnick, R., 2021b).

BATTERY

HYDROGEN

2.1.7 Wheel-Rail Interface

By collecting operational data on conditions such as temperature, wear and energy consumption, insights into the asset's performance and health can be generated (Burroughs, D., 2021b). In this sense, management of the wheel-rail interface is critical to maximizing the life of wheels and rails through

WHEEL-RAIL

preventative maintenance regimes that ensure all activities (e.g., wheel turning) offer value for money and safe operation (i.e., maintain conicity) (Vickerstaff, A., Bevan, A., and Boyacioglu, P., 2020).

FRICTION	Obviously, friction is the underlying root cause of wheel-rail degradation. The level of friction is a function of total rolling distance, effective sliding length, and sum velocity. The most dominant factor depends on the friction modifier and the working mechanism for friction stabilization. It is shown that the wear rates do not depend significantly on slip, which makes it possible to predict wear behavior. Wear rates are dependent fundamentally on the type of friction modifier used (Oomen, M. A., Bosman, R., and Lugt, P. M., 2017). Friction modifiers can effectively reduce the wheel-rail adhesion level and change the negative friction characteristic to positive. The stick-slip oscillation, which occurs in the dry clean wheel-rail contact condition, can be effectively eliminated with the application of the friction modifiers (Zhang, P., Yang, Z., Moraal, J., Dollevoet, R., Zoeteman, A., and Li, Z., 2022).
WEAR RATE	
ADHESION	
VIBRATION OPTICAL	There are several sensing technologies such as vibration and optical measurements that can be used to monitor the degree of degradation of the wheel-rail interface. They are described as follows:

Vibration Analysis High-frequency noise (i.e., above 10kHz) is generated mainly by the outside leading wheel of each bogie (Kawaguchi, T., Sueki, T., Kitagawa, T., Nishimura, M., and Abe, H., 2019). Embedded sensors installed in the wheelset provide an efficient opportunity to detect early defects in the rolling surface such as wheel flats or localized Rolling Contact Fatigue (RCF) damage. To this purpose, the Root Mean Square and the Crest Factor of the vibration signature captured at different running speeds are computed (Jarillo, J. M., Moreno, J., Alfi, S., *et al.*, 2021).

Optical Measurement Predictive wheelset maintenance using an optical measurement system with the integrated Calipri principle (i.e., a laser light section technology) for wheelset and rail maintenance allows non-contact measurement, removing all external variables such as environment conditions. The system, which reduces measurement time dramatically, provides various parameters such as the wheel profile, diameter, back-to-back distance and brake disc thickness (Burroughs, D., 2021b).

Wheel Condition Monitor Strain gauge-based system that provides information on wheel tread and loading conditions to help improve wheel life, bogie maintenance and safety. The range of detectable problems

include: wheel flats, spalls, shellings, out-of-roundness, high wheel impact loads, vehicle-axle overloading, poor vehicle/bogie loading (imbalance), and wheel unloading (Man, T., 2018).

Wheel Impact Load Detector Provides a system to detect wheel damage by measuring the peak vertical track forces and maximum dynamic ratio. This approach has proven useful in predicting the level of damage from high mean dynamic ratio recordings which are sometimes not picked up by wheel measuring tools (Groom, S., Doshi-Keeble, F., and Williams, P., 2022).

Wheel Load The diagonal wheel load imbalance is an appropriate metric to use for the detection of arbitrary vehicle defects at a bogie or vehicle level which could give rise to reduced derailment resistance. A statistical analysis of wheel load data can be used to identify anomalous characteristics which could be symptoms of defective or degraded suspensions (Shackleton, P., Sztrauch, K., Eickhoff, B., and Bevan, A., 2022). A fusion method to associate the collected samples to their positions over the wheel circumferential coordinate is useful to detect the wheel defect (Alemi, A., Corman, F., Pang, Y., and Lodewijks, G., 2019).

Profile Measurement and Stability Approach to analyze vehicle behavior at high speed (up to 250km/h) by combining a wheel profile measurement (photo laser) with a stability measurement (vertical and lateral forces) and using their joint data (Mittermayr, P., Schmid, R., Zottl, W., Betterle, E., Occioni, G., 2019). Integrating the two independent systems allows for gaining insight from the correlation between profile shape and running behavior and the associated load collectives.

Finally, the track condition also impacts the wheelset degradation. As the wheel hardness increases, the rail wear remains the same in the straight section, but as the curve radius becomes smaller, the rail wear increases again (Trausmuth, A., Schmid, R., Dinhobl, G., and Badisch, E., 2022). Similarly, rail condition in terms of wear and corrugation can be captured from sensors such as accelerometers mounted on the axleboxes (Oberhuber, H., Neuhold, J., Orta Roca, J., Brandl, D., and Schönhuber, B., 2021), and track that is not fit for purpose can also lead to RCF (Burroughs, D., 2021a). Thus, predictive maintenance is seen as a means to protect infrastructure against these expected factors, in addition to external causes such as climate change. The use of real-time sensors and simulation with forecasting capabilities helps in decision-making (Burroughs, D., 2021b).

TRACK

ACCELEROMETER

2.1.8 Brake

- While the rail industry has devoted a significant amount of research to better understand the causes of low adhesion and suitable mitigation, there remain gaps in knowledge to overcome this challenge. Continuing optimization of the wheel-rail interface and the achievement of reliable braking are challenges faced by railway networks around the world, because their improvement offers capacity increase (Englbrecht, M., 2022). Specifically, deceleration control, wheel slide protection, and smart sanding improve braking (Altman, B., and Odetunde, S., 2022). In turn, better brakes have an impact on carbon pads and wheel treads.
- BRAKING**
- LEAKAGE** Air leakage in braking pipes is a commonly encountered mechanical defect on trains. A severe air leakage will lead to braking issues and therefore decrease the reliability and cause train delays or stranding. Air leakage causes a failure when the compressor idle time is shorter than the compressor run time, that is, when the speed of air consumption is faster than air generation (Lee, W.-J., 2017). Modeling the pneumatic brake system to estimate the pressure along the brake line in real time and understanding its behavior during operation is crucial to identify the causes of technical problems and to improve driving techniques (Teodoro, I. P., Ribeiro, D. F., Botari, T., Martins, T. S., and Santos, A. A., 2019).
- COMPRESSOR**
- VIBRATION** Finally, the root problem of rail noise is the braking technology used, which affects the wheels' surface and increases its roughness, resulting in more rolling noise (Burroughs, D., 2018). In turn, the severe vibration and high stress state of the brake disc could cause it to crack in the region near the bolts (Wang, Z., Mo, J., Gebreyohanes, M. Y., Wang, K., Wang, J., and Zhou, Z., 2022).

2.1.9 Bearings

- AXLE** The axle bearing is a heavy-duty safety-critical railway element (Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., and Wang, W., 2015). It bears the weight of the train, minimizes the friction with the rotating axle, and its failure in service might cause derailment. Therefore, its maintenance is of utmost importance to guarantee the availability of the fleet.
- GREASE** While intrusive monitoring practices involving the extraction of a sample of grease and the ensuing wear particle analysis dispel all doubts about the actual degradation condition of a bearing, usually these assets are inspected on a frequent basis using other less intrusive methods. In this sense, the physical variables that lead these inspections are the temperature and the vibration.

Temperature

Hot Axle Box Detectors are used as a safety component in the railways, but identify faulty bearings too late to be useful for maintenance scheduling (Entezami, M., Roberts, C., Weston, P., Stewart, E., Amini, A., and Papaelias, M., 2020). The existing bearing temperature fault detection and early warning systems have a high false alarm rate and insufficient early warning ability. Thus, the bearing temperature data is oftentimes fused with other spatio-temporal features (including exterior temperature, train speed, etc.), and time series anomaly detection methods are then utilized to conduct the diagnosis. Results indicate the combined methods can effectively evaluate the bearing condition and provide supportive information for condition-based maintenance (Si, J., Shi, H., and Yang, J., 2022; Liu, Y. Z., Zou, Y. S., Wu, Y., Zhang, H. Y., and Ding, G. F., 2022; Garrido Martínez-Llop, P., Sanz Bobi, J. de D., and Huera Plaza, A., 2022).

TEMPERATURE

Vibration

Vibration analysis is a strategy that monitors the acoustic signature of each axle at line speed and identifies the presence of rolling surface defects in bearings while automatically ranking their severity: a bearing fault excites a structural response and this vibration radiates from the housing as sound pressure waves (Man, T., 2018). Among the different radiated energy bands, Ultrasound Acoustic Emissions ($>100\text{kHz}$) can detect early stage defects, including lubrication contamination, earlier than the other sensing technologies (Entezami, M., Roberts, C., Weston, P., Stewart, E., Amini, A., and Papaelias, M., 2020). For example, accelerometers typically provide bandwidths an order of magnitude lower. However, the start of a degeneration, i.e., the incipient point of failure, can still be determined up to 6-7 months before the asset is replaced (Barrow, K., 2018a).

VIBRATION

Experimentally, a bearing test rig for Very High Speed Trains has confirmed the feasibility of a CBM approach on several bearing damage such pitting and oxidation on rollers, inner and outer races, for different environmental conditions. Envelope Analysis and Spectral Kurtosis were the most effective and robust techniques, both at constant speed and in slow transients, such as the ones that characterize the rail environment (Pennacchi, P., Bruni, S., Chatterton, S., Borghesani, P., Ricci, R., Marinis, D., Didonato, A., and Unger-Weber, F., 2011). Other useful signal processing indicators and techniques include the Time-Synchronous Averaging for filtering (at axle angular speed), statistical features (e.g., root mean square, peak-to-peak, kurtosis, crest factor, skewness), and specific bearing algo-

ACCELEROMETER

rhythms (e.g., ball energy, cage energy, inner race energy, and outer race energy) (Zhu, J., Nostrand, T., Spiegel, C., and Morton, B., 2014). The results show that the longitudinal vibration features (i.e., the peak-to-peak value) are more sensitive for inner race fault identification, while the vertical vibration features (i.e., skewness value) are more suitable for outer race fault identification (Wang, J., Yang, J., Bai, Y., Zhao, Y., He, Y., and Yao, D., 2021).

2.1.10 Pantograph–Catenary Interface

PANTOGRAPH Railway pantographs are used around the world for collecting electrical energy to power railway vehicles from the overhead catenary (Xin, T., Roberts, C., Weston, P., and Stewart, E., 2020). Faults in the pantograph system degrade the quality of the contact with the catenary and reduce the reliability of railway operations. To track the degradation of this asset, a Contact Line Monitoring system captures video data with cameras installed on commercial train cars and uses AI technology to analyze the stream of images. This system has enabled significant reductions in the number of personnel needed to perform inspections, and has also helped to achieve low-cost operations. Additionally, the image analysis using AI technology has improved the precision of the inspections (Matsumoto, T., Nishidouzono, K., Fukaya, F., Koga, S., Nakamura, H., and Kameda, M., 2022).

CATENARY Condition detection and evaluation of the pantograph–catenary electrical contact is mainly led by three technologies: ultrasounds, image recognition, and spectral diagnosis (Wu, G., Dong, K., Xu, Z. *et al.*, 2022). However, additional sensing inputs such as the acceleration on the bow suspension of the pantograph is helpful to detect the shocks and trigger the recordings (Ben Taleb Ali, M., Schrevre, T., Pedron, A., Blanvillain, G., and Auditeau, G., 2022). Common defect types that are detectable using this approach include contact wire splice, neutral section, section insulator, catenary obstacle, etc.

SLIDING CONTACT The sliding contact between the pantograph and catenary is what allows the correct operation of electric trains. This contact induces wear on the components involved in the process, namely the carbon strip on the pantograph and the metal wire on the catenary. Heuristic wear models (as complex parametric functions) can be used together with statistical data of the current, contact force, and line speed retrieved from field measurements to predict the wear when operations are carried on along a railway line (Derosa, S., Nåvik, P., Collina, A., Bucca, G., and Rönnquist, A., 2020). The lateral speed does not affect the mechanical wear and electrical wear associated with the arcs that commonly occur. However, it affects the electrical term

CARBON STRIP

associated with the Joule effect (Derosa, S., N avik, P., Collina, A., Bucca, G., and R onnquist, A., 2021). Lastly, multi-body approaches model the pantograph with detail and can accommodate the non-linear characteristics of the real system and include external loads that act on the pantograph during operation, e.g., aerodynamic loads, which are especially important when studying catenary gradients (Rebelo, J., Pombo, J., Antunes, P., Santos, J., Magalh aes, H., and Ambr osio, J., 2022).

Such physical-numerical hybrid simulation can represent the dynamic interaction between the pantograph and the overhead line, and phenomena due to the span-passing frequency such as the uplift of the contact wire at the support point can be evaluated. Furthermore, the effect of friction due to sliding between contact strips and contact wire is also represented in these simulated data (Kobayashi, S., Koyama, T., and Harada, S., 2022). However, the capability of these numerical tools has been, in general, limited to pantograph-catenary dynamic analyses set in a single straight railway track (Antunes, P., Pombo, J., Ambr osio, J., Rebelo, J., Santos, J., 2022).

All this modelling effort has also led to focus on reliability approaches for maintaining the pantograph-catenary interface. First, the reliability of the key parts of the system are modeled using Weibull distribution. Second, a reliability margin is proposed to expand the maintenance time from point to interval, and the reliability margin is optimized to minimize the maintenance cost. Then, a preventive opportunistic maintenance schedule can be arranged on the basis of the optimal reliability margin. This method can reduce the number of maintenance schedules and can effectively save the maintenance cost (Cheng, H., Cao, Y., Wang, J., Zhang, W., and Zeng, H., 2020).

OVERHEAD LINE

FRICTION

RELIABILITY

2.2 Technologies Applied to Product Development

Regarding the application of AI in the paradigm of maintenance optimization for the Industry 4.0 Pinciroli, L., Baraldi, P., and Zio, E. (2023), the main scientific research efforts published in the surveyed papers have been seen in the subdomain of rail maintenance and inspection (Tang, R., De Donato, L., Bešinović, N., Flammini, F., Goverde, R. M. P., Lin, Z., Liu, R., Tang, T., Vittorini, V., and Wang, Z., 2022; Bešinović, N., De Donato, L., Flammini, F., *et al.*, 2022). What follows is a deeper analysis into some of the technologies that have enabled the application of AI to railway predictive maintenance.

2.2.1 Smart Sensors

INTERNET-OF-THINGS	<p>The Industrial Internet-of-Things (IoT) has emerged as one of the leading technologies to deploy the remote condition monitoring of machines (Boyes, H., Hallaq, B., Cunningham, J., and Watson, T., 2018; Compare, M., Baraldi, P., and Zio, E., 2019), especially when such machines are transportation assets that move around the territory. With the range of sensors and IoT solutions on the market growing exponentially, operators and infrastructure managers are gathering more data than ever, which signals a change in the culture. However, despite of the successful results from the proof of concepts, many cases failed or were dropped because the business case was not obvious (Burroughs, D., 2019a). In this sense, the failure to start a pilot or to go beyond its initial deployment, maybe due to IT arrogance, complications, budget, focus, etc., is what thwarts its potential rollout in the market (Miciek, R., 2019).</p>
REMOTE	<p>Remote monitoring equipment provides a centralized system for fleet diagnostics and supports the shift to predictive maintenance (Smith, K., 2018). However, as the volume of data increases and ever more devices become connected to the IoT, centralized processing of data could become impractical due to long cycle times and high consumption of computing resources, which conflict with the real-time response requirements of fault diagnosis (Zhang, K., Huang, W., Hou, X., Xu, J., Su, R., and Xu, H., 2021). Therefore, it may be necessary to transfer processing capability out into the field, and edge computing is currently a focus for IoT solutions (Barrow, K., 2018a). In this approach, various sensors are installed on the asset, and upon acquiring the data and extracting the information, they create a baseline pattern (Aimar, M., and Somà, A., 2018), and prescriptions are eventually pushed to the cloud (Burroughs, D., 2021b).</p>
EDGE COMPUTING	<p>Alstom has been developing different IoT network products for over 10 years with a focus on vibration analysis through accelerometers (Trilla, A., and Gratacòs, P., 2013, 2016; Trilla, A., Gratacòs, P., Guinart, D., Alessi, A. and Lamoureux, B., 2016; Trilla, A., Janjua, F., and Bermejo, S., 2019). Some of the developed solutions focus on the needs of the shop floor for conducting inspections at the depot. Others focus on the needs of the remote maintainer, and carry out synthetic indices from vibration measurements while a gateway acquires correlated GPS and odometry information. These products pave the way to the future development of a completely wireless system able to perform condition monitoring of both the vehicle and the infrastructure (Zanelli, F., Sabbioni, E., Carnevale, M., <i>et al.</i>, 2023). Finally, the joint operation of the IoT and AI are also offering new insights into how cities function and how services can be optimized to meet the everyday</p>
ACCELEROMETER	
VIBRATION	

needs of urban dwellers (Barrow, K., 2019b).

2.2.2 Machine Vision Inspection

The technology provided by machine vision software provides the tools for rail companies to rethink their approach to reliability engineering (Kilian, K., Kilian, M., Mazur, V., and Phelan, J., 2016). Specifically, Machine Learning techniques are at the foundation of Computer Vision. In addition to helping increase the capacity of the line, e.g., reducing the headway between trains by 4 to 8 minutes, it can enable railways to monitor rolling stock and infrastructure, and to target maintenance and repair activities (Romanchikov, A., and Smith, K., 2022). Additionally, the real-time detection of interesting items (e.g., track, people, signs, animals...) can improve the driver situational awareness and train protection (Burroughs, D., Wust, D., and Wust, J., 2023).

COMPUTER VISION

Illustrative applications of Machine Vision Inspection for predictive maintenance include different solutions for the brakes and the wheels:

- A Brake Inspection Monitor diagnoses the brakes of trains passing the site. It measures the remaining material in each brake pad and then calculates a replacement window based on the historical wear rate. Important parameters that are captured include: pad thickness, pad wear rates, sticking brakes, presence of the brake key, and identification of missing brake pads (Man, T., 2018).
- A Wheel Profile Monitor measures the wheel profile of trains that pass the installation site at steady operating speeds. The system firstly captures images of the wheels and processes them with advanced machine vision algorithms to measure key wheel parameters, including: flange height and width, tread hollowing, back-to-back dimension, inner and outer rim thickness, wheel diameter and differential, and wheel profile trace (Man, T., 2018). More innovative systems use digital high-speed imaging along with 3D laser scanning for obtaining these parameters (Burroughs, D., 2021b).

Alstom has developed a train monitoring system that is aimed at optimizing the maintenance of brake pads, pantograph carbon strips, and wheelsets, through the deployment of the PHM methodology and its associated techniques. It integrates a series of acquisition subsystems with lasers and 3D cameras that capture the related measures as a train traverses its portal. Then, it automatically conducts the processing and analysis of the

collected data, and finally it triggers alarms and issues reports to the maintenance staff (Lortie, M., and Holmes, E., 2014; Trilla, A., and Cabré, X., 2018; Trilla, A., Dersin, P., and Cabré, X., 2018).

2.2.3 Technical Language Processing

TEXT MINING In the railway industry, a significant amount of data is stored in the textual format. The advanced development of Natural Language Processing (NLP) and text mining techniques enable automatic knowledge extraction and discovery from such documents. Text-based research and analysis are bound to have a meaningful impact on almost every aspect of the railway sector (Dong, K., Romanov, I., McLellan, C., and Esen, A. F., 2022). Research shows that over 82% of the NLP-related papers were published in the last 5 years, indicating a growing research interest in this field.

SAFETY Railway safety is the most studied topic in text-based railway research with over 28% of published papers, and is closely followed by fault diagnosis with 25% of published articles (Dong, K., Romanov, I., McLellan, C., and Esen, A. F., 2022). Communication errors in railway systems could pose a serious threat to safety (Nakamura, R., 2019). On average, the cause of these errors is found in the ambiguity of the language used to express the information. Overall, NLP has the potential for analyzing accident data (Valcamonico, D., Baraldi, P., Amigoni, F., and Zio, E., 2022; Heidarysafa, M., Kowsari, K., Barnes, L., and Brown, D., 2018; Song, B., Zhang, Z., Qin, Y., Liu, Y., and Hu, H., 2022), and assist the risk and incident analysis experts to study causal relationships on failures towards the overall safety in the rail industry (Liu, J., Schmid, F., Li, K., and Zheng, W., 2021; Syeda, K. N., Shirazi, S. N., Naqvi, S. A. A., Parkinson, H. J., and Bamford, G., 2019). The frequency, distribution, and co-occurrence of text concepts form unsupervised patterns that can provide useful indicators for investigations, and assist incident experts in establishing root cause analysis using relevant supporting information (Farzad, A., and Gulliver, A., 2020). However, they require processing approaches that are different from parametric sensor data (Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020).

NLP provides an effective approach for improving the collection and analysis of text-based maintenance data, and eventually enable accurate decision-making (Brundage, M. P., Weiss, B. A., and Pellegrino, J., 2020). NLP can be applied to text entry fields of maintenance records to guarantee the data quality before doing any statistical analysis or making any decision (Stenström, C., Al-Jumaili, M., and Parida, A., 2015). To process large amounts of unstructured text information about railway equipment faults in

the form of natural language, topic models have been used to extract the semantic features of the text, and text classifiers have been used to construct a signal equipment fault diagnostic model (Shi, L., Zhu, Y., Zhang, Y., and Su, Z., 2021).

TOPIC MODEL

Finally, the bleeding edge in industrial NLP research is called Technical Language Processing (TLP), which presents a holistic, domain-driven approach, to use NLP in a technical engineering setting (Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S., 2021). In TLP, maintenance documents like work orders are relatively small in size and contain misspellings, domain-specific jargon, abbreviations, and non-standard sentence structure.

TECHNICAL LANGUAGE

WORK ORDER

2.2.4 System Log Analytics

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 (Sammouri, W., Côme, E., Oukhellou, L., Aknin, P., and Fonlladosa, C.-E., 2014). In this sense, fuzzy classification enables the use of linguistic variables for the definition of the time intervals in which the failures are predicted to occur, which provides a more intuitive way to handle the predictions by the users, and increases the acceptance of the proposed approach (Fink, O., Zio, E., and Weidmann, U., 2015c). Also, when only discrete-event data are available, tackling a regression problem as a classification by fixing a time interval was found to be sufficiently precise for the operators to be able to anticipate the prediction and schedule a pertinent maintenance task prior to the occurrence of the event (Fink, O., Zio, E., and Weidmann, U., 2015a,b; Lee, W.-J., 2017).

NOMINAL VARIABLE

EVENT DATA

Finally, identifying frequent item-sets is also a popular data-mining task to discover association rules in the events (Shabtay, L., Fournier-Viger, P., Yaari, R., and Dattner, I., 2021). Furthermore, in order to tackle the combinatorial explosion problem, on-the-fly and incremental techniques for fault diagnosis of discrete event systems have also been explored (Liu, B., Ghazel, M., and Toguyéni, A., 2018).

DEEP LEARNING-BASED PHM

Every time I find a system that isn't a gradient search, I insert a smooth curve; then I can cast it in the general form of a gradient search. In fact, that's about all there is to neural networks.
– David E. Rumelhart (1993)

THERE are two fundamental approaches to tackle PHM objectives: model-based and data-driven. Moreover, these are not mutually exclusive. Hybrid physical/data-driven approaches such as the digital twin utilize virtual representations of some aspect of a system to provide quantifiable benefit (Moyné, J., Balta, E. C., Kovalenko, I., Faris, J., Barton, K., and Tilbury, D. M., 2020). However, there are no context-independent or usage-independent reasons to favor one learning method over another. Methods that are effective for forecasting risk and informing maintenance decisions for individual components do not readily scale to sub-system or system level insights. A holistic modeling approach is needed, one that incorporates available structural and physical knowledge and naturally handles the complexities of actively fielded and maintained assets (Miller, K., and Dubrawski, A., 2019). System conceptualization modeling and verification, dynamic and automatic pattern recognition and environment adaptation, and sequential decision optimization are key points to increase the chances of success (Hu, Y., Miao, X., Si, Y., Pan, E., and Zio, E., 2022).

DIGITAL TWIN

The data-driven techniques on which this dissertation is focused are

DATA-DRIVEN

based on collecting experimental data and extracting meaningful features to determine if the system is normal (i.e., the healthy condition) or are there any symptoms of failure. If the latter is true, the failure must be classified and categorized to identify the fault and determine its severity (Khan, S., and Yairi, T., 2018). Depending on the way this knowledge is learned, there are different ways to approach the solution. In the supervised case, the operating user shall have complete data of all failures modes and expected behaviors. In the semi-supervised case, the user has access to limited data, e.g., often only healthy data is available. In the worst-case scenario, i.e., the unsupervised case, the system is already operating and there is no knowledge about its condition (Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020).

SUPERVISED

UNSUPERVISED

This chapter is organized as follows: Section 3.1 provides a review of the Deep Learning fundamentals that triggered the recent wave of neural network research, along with its challenges and open questions. Then, Section 3.2 describes the specific techniques used in product development that address the former topics.

3.1 Main Topics in the Deep Learning Research Community

DEEP LEARNING

Deep Learning (DL) has gained increasing attention due to its advantages in data classification and feature extraction problems. It is an evolving research area with diverse application domains. Hence, it has the potential to increase overall system resilience or cost benefits for maintenance, repair, and overhaul activities (Khan, S., and Yairi, T., 2018).

3.1.1 Breakthroughs in Neural Networks

NEURAL NETWORKS

In the literature, there are several excellent reviews by leading scientists in the DL field as the latest generation of neural networks (Schmidhuber, J., 2015; LeCun, Y. and Bengio, Y., and Hinton, G. E., 2015; Raghu, M., and Schmidt, E., 2020). This section provides a minimal set of ideas compiling most of the basic knowledge necessary to understand modern DL research, in historical order (Wang, H., and Raj, B., 2017; Britz, D., 2020).

Multilayer Perceptron and Backpropagation (1986)

MULTILAYER PERCEPTRON

A Multilayer Perceptron (MLP) is basically a shallow state-less feed-forward network with *at least* 3 layers: an input layer I , a hidden layer

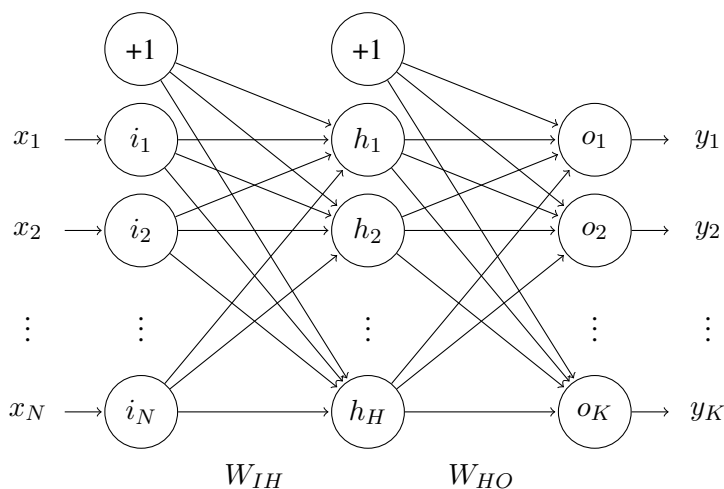


Figure 3.1 Multilayer Perceptron network where $O = g(I \cdot W_{IH}) \cdot W_{HO}$. The non-linear function g is inherent in the hidden layer. In this diagram, the bias terms “+1” are made explicit.

H , and an output layer O . These layers are pairwise connected with two dense networks W_{IH} and W_{HO} . Additionally, the intermediate hidden layer equips a non-linear activation function g that enables the MLP to represent any continuous function given enough model expressiveness, i.e., the number of nodes H . This is known as the Universal Approximation Theorem (Cybenko, G., 1989; Hornik, K., 1989). Finally, the MLP is suitable for tackling hetero-association problems, i.e., relating an arbitrary input X (with vector length N) to an arbitrary output Y (with vector length K), see Figure 3.1.

Traditionally, MLP’s were trained using backpropagation gradient descent where the weights were updated for each layer as a function of the derivative of the previous layer (Rumelhart, D. E., Hinton, G. E., and Williams, R. J., 1986). However, there were limitations to this approach when the depth of the network was increased because these gradients vanished. Today, DL provides a series of methods that overcome the vanishing gradient problem and therefore enable training larger neural networks (Khan, S., and Yairi, T., 2018).

Autoencoders, Embeddings, and Weight Pre-training (2006/2009)

The Autoencoder is a particular neural architecture that inherently learns to replicate data through a compressed representation in the middle “bot-

AUTOENCODER

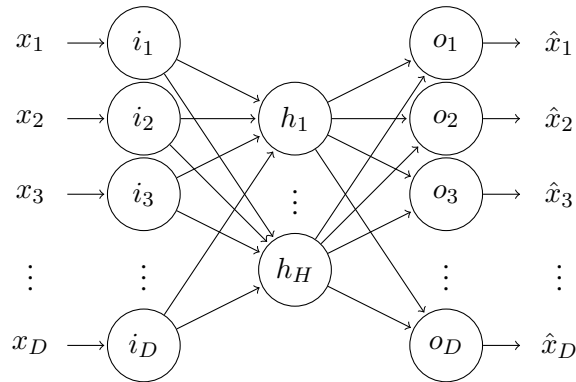


Figure 3.2 Autoencoder architecture, where D is the data dimensionality and H is the size of the hidden layer, which defines the representational capacity of the network. The compressive encoding function of the model is ensured as long as $H < D$.

EMBEDDING tleneck” layer, i.e., the embedding. This way, it extracts the most useful/relevant features from the data in an unsupervised auto-associative manner (Kramer, M. A., 1992; Stone, V. M., 2008). Given that the task of auto-association is easier than hetero-association, Autoencoders were the first technique that showed success in building the early deep neural designs (Hinton, G. E., and Salakhutdinov, R. R., 2006).

RECONSTRUCT The architecture of an Autoencoder shows a convergent structure from its input dimensionality D into H at half of its depth, and then it diverges back to D toward its output, see Figure 3.2. The Autoencoder is trained to encode the input into some lower-dimensional representation so that it may thereafter be reconstructed. As a result, the network learns a compressed distributed representation of the data that captures its main factors of variation (Bengio, Y., 2009).

PRE-TRAINING Finally, the concept of pre-training refers to first training a model to perform a given task, and then reusing the learned embedding through its parameters as an initialization to learn a new model on a related task. This head-start to the optimization procedure is what was initially used to create the first deep neural architectures using on stacks of Autoencoders.

Xavier Initialization (2010)

When the weights of a neural network layer are initialized with normally-distributed values, it is easy for them to explode or vanish, therefore preventing training. Assuming the values from the previous layer follow Gaussian distributions, their variances accumulate, and thus they should be scaled

down proportionally to the number of inputs to keep them bounded. The same criterion holds in the reverse direction (i.e., with the number of outputs). Xavier Initialization takes these fan-in and fan-out considerations into account to randomly initialize the weights of a neural network (Glorot, X., and Bengio. Y, 2010).

INITIALIZATION

Rectified Linear Unit (2011)

Traditional neural networks used sigmoids for their intermediate activations. Sigmoids (most commonly the logistic and hyperbolic tangent functions) have the advantages of being differentiable and having a bounded output. However, their derivatives decay quickly away from zero, and as more layers are stacked, the gradients disappear. This is known as the vanishing gradient problem and is one of the reasons that networks were difficult to scale depthwise.

VANISHING GRADIENT

Rectified Linear Units (ReLU) helped solve the vanishing gradient problem and paved the way for deeper networks (Glorot, X., Bordes, A., and Bengio. Y, 2011). The derivative of a ReLU is a step function, and thus prevents the positive gradients from disappearing. However, ReLU still have some flaws: they are non-differentiable at zero, they can grow unbounded, and neurons could become inactive due to saturation.

RECTIFIED

Dropout (2012)

One of the most prominent reasons for causing overfitting is co-adaptation, which means that some neurons are highly dependent on others. However, overfitting is greatly reduced by randomly omitting (i.e., dropping), half of the units on each training case (Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R., 2012; Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., 2014). The use of Dropout, which introduces a random deactivation of units, has become a crucial component for improving the generalization ability of all kinds of neural models.

OVERFITTING

DROPOUT

Convolutional Neural Networks (2012)

The Convolutional Neural Network (CNN) is a particular deep neural architecture that uses a convolution operation in place of the general matrix multiplication. This process naturally involves learning a set of filters (or kernels) that detect useful local patterns in the input signals. This fact turned the conventional manual feature extraction design into an automated pro-

CONVOLUTIONAL

LOCAL PATTERN

FEATURE EXTRACTION

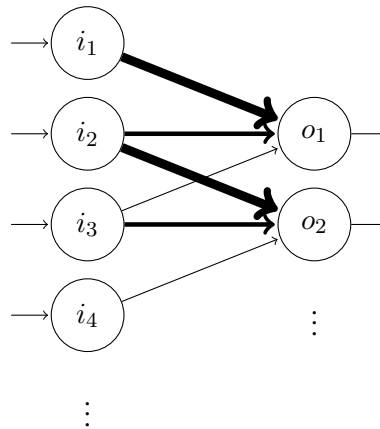


Figure 3.3 Convolutional layer for an input vector $I = (i_1, i_2, \dots)$, a 1D second-order filter $W = (w_1, w_2, w_3)$, and an output vector $O = (o_1, o_2, \dots)$, which clearly shows that $O = I * W$. Edge thickness indicates parameter reuse.

cess, which is its primary advantage. Additionally, since the weights of the filters are shared and reused throughout this sparsely connected architecture, there are less parameters to be learned, and this makes the CNN less prone to overfit the data (Duda, R. O., Hart, P. E., and Stork, D. G., 2001). For example, Figure 3.3 shows the architecture of a convolutional layer.

The linear function that the CNN layer implements, which is the convolution, is sometimes defined as the cross-correlation function in the literature. For real-valued functions of a continuous or discrete variable, convolution differs from cross-correlation only in that the filter (or the input signal) is reflected in the computation. From a pragmatic standpoint, where the objective is to extract salient data characteristics, this nuance is not relevant, though. The features that the CNN extracts are simply good-enough for all tasks, including classification (i.e., diagnosis) and regression (i.e., prognosis) objectives (Jernelv, I. L., Hjelme, D. R., Matsuura, Y., and Aksnes, A., 2020).

The revolutionary architecture comprising a sequence of convolutional layers, ReLU nonlinearity, and max-pooling (i.e., subsampling), became the accepted standard for Computer Vision applications after performing significantly better than previous methods at classifying images (Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012).

Distributed Representations (2013)

In a distributed representation, each entity is represented by a pattern of

activity over many computing elements, and each computing element is involved in representing many different entities. This gives rise to dense vector spaces (in contrast to sparse local representations where each entity is represented by one computing element). A notable illustration of this concept is the word embedding, which became the dominant way to encode text for DL Natural Language Processing models (Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013; Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., 2013; Mikolov, T., Yih, W.-t., and Zweig, G., 2013). This concept is based on the idea that words which appear in similar contexts likely have similar meanings, and thus can be used to embed words into vectors to be used downstream in other models.

DENSE VECTOR SPACE

EMBEDDING

Deep Reinforcement Learning (2013)

Reinforcement Learning differs from Supervised Learning, i.e., example driven learning, in that an agent must learn to maximize the sum of rewards over multiple time steps instead of just predicting an outcome. In this scenario, the agent interacts directly with the environment (i.e., an intervention) and each action affects the next. Thus, the training data is not independent and identically distributed, which makes the training of a straightforward DL model unstable.

REINFORCEMENT LEARNING

The objective for training the value-based reward function is derived from the Bellman equation, which decomposes it into the current reward plus the maximum (discounted) future reward of the next state. The breakthrough application of this paradigm was attained by playing computer games from raw pixel inputs (Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M., 2013).

REWARD

Encoder-Decoder Networks with Attention (2014)

Traditionally, learning sequences was tackled through a recurrent network such as a Long Short-Term Memory (LSTM), which maintains a memory of previously processed inputs (Sutskever, I., Vinyals, O., and Le, Q. V., 2014). However, these recurrent models often had difficulty dealing with dependencies over long time horizons, and would “forget” earlier inputs because their gradients needed to propagate through many time steps. Today, the attention mechanism helps alleviate the problem by unfolding the sequence in time and introducing shortcut connections, thus giving the network an option to adaptively review earlier time steps (Bahdanau, D., Cho, K., and Bengio, Y., 2015).

RECURRENT

ATTENTION

Adam Optimizer (2014)

LOSS FUNCTION Neural networks are generally trained by minimizing a loss function using a gradient descent optimizer. However, many of these optimizers contain many tunable hyperparameters, and finding the right settings for a specific problem not only reduces training time but can also lead to better results due to finding a better local minimum of the loss function. The Adam optimizer was proposed to use the first and second moments of the gradients to automatically adapt the learning rate for each parameter separately (Kingma, D. P., and Ba, J. L., 2015). The result turned out to be quite robust and less sensitive to hyperparameter choices.

Generative Adversarial Networks (2014)

GENERATIVE MODEL The goal of generative models is to create realistically-looking data samples. The basic idea behind Generative Adversarial Networks (GAN) is to train two networks in tandem: a generator and a discriminator. The goal of the generator is to produce samples that fool the discriminator, which is trained to distinguish between real and synthetic (i.e., artificially generated) data (Radford, A., Metz, L., and Chintala, S., 2016). Relying on a mini-max game between the two networks, GANs are able to model complex high dimensional distributions (Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., 2014).

Residual Networks (2015)

RESIDUAL SHORTCUT CONNECTION Traditionally, training deep networks has been a challenging optimization problem due to the vanishing gradients. Residual Networks (ResNet) provided a workaround by using identity shortcut connections between the input and the output of different layers that help the gradients flow (He, K., Zhang, X., Ren, S., and Sun, J., 2015). This trick enabled learning very deep neural architectures.

Batch Normalization (2015)

NORMALIZING Training deep neural networks is complicated due to the fact that the distribution of each layer's inputs changes during training, as the parameters of the previous layers change. This problem is known as internal covariate shift, and its solution involves normalizing the layer inputs by tracking the statistics during training in batches, and using them to scale activations to zero mean and unit variance (Ioffe, S., and Szegedy, C., 2015).

Transformers (2017)

The traditional problems that recurrent networks had for learning sequences were definitively solved by Transformers, which unfolded the recurrence and introduced multiple feed-forward self-attention layers, processing all inputs in parallel, and producing relatively short paths between inputs and outputs (Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. M., Kaiser, L., and Polosukhin, I., 2017), making them suitable for gradient descent optimization. Additionally, Transformers had to use positional encodings to inform about the order of the data, which was implicit in the recurrence. Transformers have since then become the standard architecture for the vast majority of NLP and other sequence tasks, and have even made their way into architectures for Computer Vision.

TRANSFORMER

Neural Architecture Search (2017/2018)

Neural Architecture Search (NAS) has become common practice in the field for maximizing the performance of neural networks. Instead of manually designing architectures, NAS allows this tedious process to be automated (Zoph, B., and Le, Q., 2017). To accomplish this task, different criteria can be considered: search space, search strategy, and evaluation strategy (Elsken, T., Metzen, J. H., and Hutter, F., 2019; Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., and Wang, X., 2020).

ARCHITECTURE SEARCH

Bidirectional Contextual Embedding (2018)

Bidirectional Contextual Embedding (BERT) was the latest of such pre-training developments that gave birth to the first deep neural architectures in 2006 (Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., 2018). In the application context of NLP, BERT learned a language model that was pre-trained on predicting masked (intentionally removed) words anywhere in the sentence (i.e., bidirectionality), and whether if two sentences were likely to follow each other. This unsupervised pre-trained model, which learned some general properties about language, could then be fine-tuned to solve supervised tasks.

BIDIRECTIONALITY

Transfer Learning (2018)

Transfer Learning focuses on reusing a model (and thus its inherently learned knowledge) trained on one problem to a different but related problem or environment. Transfer Learning is classified into four categories: instances-based (weight instances in the source domain), mapping-based

TRANSFER LEARNING

(new space with better similarity), network-based (pre-training in the source domain and fine-tuning in the target domain), and adversarial-based (find features suitable for the two domains). Recent research focuses on transferring knowledge using unsupervised or semi-supervised learning for the increased availability of data (Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C., 2018).

Double Descent (2019)

BIAS-VARIANCE Traditionally, model expressiveness follows the bias-variance tradeoff: it must match the structural complexity of the data to avoid underfitting (too few parameters) and overfitting (too many parameters). In practice, however, DL models are often overparameterized and yet exhibit a good performance (Belkin, M., Hsu, D., Ma, S., and Mandal, S., 2019; Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I., 2019).
 DOUBLE DESCENT Double Descent posits that gradient descent is more likely to find these smoother zero-training-error networks, which generalize well despite being overparameterized.

The Lottery Ticket Hypothesis (2019)

LOTTERY TICKET The Lottery Ticket Hypothesis asserts that most credit of a performant model comes from a certain inherent subnetwork due to a lucky parameter initialization. Thus, larger models have a higher chance of having these subnetworks. This line of research focuses on pruning the irrelevant weights and using the ones that remain as pre-training, which yields a performance close to the original loss (Frankle, J., and Carbin, M., 2019).

Large Models, Self-Supervised Learning, and Knowledge Distillation (2019/2020 and beyond)

PARALLELIZATION The clearest trend throughout the history of DL is perhaps that of the bitter lesson (Sutton, R. S., 2019), which states that the algorithmic advances for better parallelization (enabling more model parameters) win over smarter learning techniques. As models become bigger and faster to train, techniques that can make efficient use of the huge set of unlabeled data on the web, and learn general-purpose knowledge that can transfer to other tasks, are becoming more valuable and widely adopted (Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., 2020).

Also, deploying deep models on devices with limited resources, e.g., mobile phones and embedded devices, is challenging not only because of

the high computational complexity but also for the large storage requirements. To this end, a variety of model compression and acceleration techniques have been developed. For example, knowledge distillation is an approach that effectively learns a small student model from a large teacher model (Gou, J., Yu, B., Maybank, S. J., and Tao, D., 2021).

KNOWLEDGE DISTILLATION

Finally, a new field of mathematical analysis has been developed around Deep Learning (Berner, J., Grohs, P., Kutyniok, G., and Petersen, P., 2022; Kutyniok, G., 2022). It emerged to address a series of research questions that were not answered within the classical framework of learning theory: the role of depth in deep architectures, the apparent absence of the curse of dimensionality, the optimization performance despite the non-convexity of the problem, the interpretability of the learned features, etc. As it usually happens with theorems, they are likely to spur the development of new corollaries that will find their way into new applications, thus opening up new avenues for DL improvement.

MATHEMATICAL ANALYSIS

3.1.2 Deep System Health Management

The aim of health management is to collect relevant data from various sensor sources and carry out the necessary processing including the extraction of key features, fault detection, fault diagnosis, and prognosis (Khan, S., and Yairi, T., 2018). In a nutshell: initially, fault detection uses either signal reconstruction error or a binary classifier on top of the network to detect anomalies. Then, fault diagnosis typically adds a soft-max layer to perform multi-class classification. Finally, prognosis adds a continuous regression layer to predict the remaining useful life (Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., and Wei, M., 2019).

HEALTH MANAGEMENT

Recent advances in detection include ensemble models using Convolutional layers (Selvanathan, B., Nistala, S. H., Runkana, V., Desai, S. J., and Agarwal, S., 2023; da Rosa, T. G., de Andrade Melani, A. E., Kashiwagi, F. N., de Carvalho Michalski, M. A., de Souza, G. F. M., de Oliveira Salles, G. M., and Rigoni, E., 2022), autoencoders (Brunner, S., Frischknecht-Gruber, C. M.-L., Reif, M., and Senn, C. W., 2022), GANs (Xu, M., Baraldi, P., Lu, X., and Zio, E. (2022), and LSTM networks (Hosseinpour, F., Ahmed, I., Baraldi, P., Behzad, M., Zio, E., and Lewitschnig, H., 2022; De Simone, L., Caputo, E., Cinque, M., Galli, A., Moscato, V., Russo, S., Cesaro, G., Criscuolo, V., and Giannini, G., 2023). In diagnosis, a Feature Importance layer, which is a one-to-one link with features, indicates the relevance for result interpretability (Barraza, J. F., Droguett, E. L., and Martins, M. R., 2021). Finally, regarding prognosis and specifically the estimation of the Remaining Useful Life (RUL), Convolutional layers and residual blocks are

DETECTION

DIAGNOSIS

PROGNOSIS

combined (DeVol, N., Saldana, C., and Fu, K., 2022), also with LSTM (Remadna, I., Terrissa, L. S., Ayad, S., and Zerhouni, N., 2021; Tamssaouet, F., Nguyen, K. T. P., Medjaher, K., and Orhard, M., 2021).

Based on the complexity breakdown analysis between diagnosis and prognosis objectives, the system should be able to recommend further actions according to user requirements, i.e., the generation of the advisory statement. This final phase plays an important role in adding resilience to the overall solution and regulating availability during service operation (Khan, S., and Yairi, T., 2018). The key points are summarized as follows:

- Any recommended decisions are only as good as the data that was collected to represent the current state of the system operation.
- False alarms have been identified as a major annoyance during maintenance activities.
- System models and related algorithms need to be updated from time to time in order to account for any unanticipated conditions.
- Recording and storing acquired on-field knowledge for future application developments and improvements is useful.

3.1.3 Challenges and Opportunities

Failures are to be prevented as much as possible to maximize the availability of the assets. To accomplish this goal, some approaches focus on designing for failures, while others put the attention on maintenance. With respect to the latter, PHM constitutes its paramount implementation. This section highlights the main challenges and opportunities for applying DL to PHM (Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., and Wei, M., 2019; Rezaeianjouybari, B., and Shang, Y., 2020; Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020):

Cross-domain Prediction Transfer learning aims to take advantage of the experience learned in a source problem to improve the learning of a target problem. However, complex systems cannot be directly transferred to a different system of the same fleet operated under different conditions, or to a different fleet.

Data Scarcity and Augmentation DL algorithms are known to be data-hungry, and their superior performance depends on the availability of

abundant data, which is rarely feasible in most of the PHM situations (failures are usually rare).

Industrial Data Characteristics The success of deep models is reliant on the quality and variety of the collected data. However, real-world industrial data come from various sensors and are mostly incomplete, imbalanced, unlabeled, unstructured, multi-modal, and heterogeneous.

Data Analysis Pre-processing actions range from simple normalization, standardization, and data segmentation, to more complex tasks such as labeling and dealing with incomplete data or outliers.

Model Selection Expert manual processes are error-prone and time-consuming. DL depends on a wide spectrum of hyperparameters for automation, which enlarge the search space for the optimum solution.

Interpretability and Explainability The lack of “transparency” affects the decision-making part of the PHM cycle. Interpretability techniques are roughly categorized into two categories: those that utilize a surrogate simple model, and those that incorporate explainable mechanisms in intermediate layers such as attention or physics-induced prior knowledge.

Real-Time Realization Actual industrial data come in continuous streams and their distribution characteristics are in dynamic change over time. PHM DL models need to cope with the concept of drift of continuously evolving new data within the incremental learning settings.

Benchmarking Towards a fair comparison of the time and cost of developing solutions, there is a need to build novel metrics that incorporate runtime performance, model accuracy, and robustness across various architectures and DL frameworks.

3.2 Learning Techniques for Product Research

This section makes some remarks about some of the important DL techniques that have found a deal of success to develop specific solutions for the railway PHM industry: Multilayer Perceptrons, Convolutional Networks, and Autoencoders.

While convolutions and attention are both sufficient for good performance, neither of them are necessary. An architecture based exclusively on

MLPs applied independently to local data, followed by more MLPs mixing the formerly extracted representation performs equivalently to a full DL architecture (Tolstikhin, I., Houlby, N., Kolesnikov, A., Beyer, L., *et al.*, 2021; Melas-Kyriazi, L., 2021). As a refinement strategy, there are several approaches that may also be considered: a structural re-parameterization that adds a local prior into the dense layers (Ding, X., Chen, H., Zhang, X., Han, J., and Ding, G., 2022), gating functions (Liu, H., Dai, Z., So, D. R., and Le, Q. V., 2021), and residual connections (Touvron, H., Bojanowski, P., Caron, M., *et al.*, 2021). Such well-regularized plain MLPs significantly outperform recent state-of-the-art specialized neural network architectures, and they even outperform strong traditional ML methods, such as XGBoost, which is the champion on tabular datasets (Kadra, A., Lindauer, M., Hutter, F., and Grabocka, J., 2021). Finally, the use of the Autoencoder in PHM highlights its capacity to detect anomalies (Goldthorpe, P., and Desmet, A., 2018) and to construct health indices (Trilla, A., Janjua, F., and Bermejo, S., 2019), among other specific applications to mechanical components such as bearings, turbines, etc. (Khan, S., and Yairi, T., 2018).

STATE OF THE ART

*The best engineer I ever knew never had an original idea in his life.
All he'd do is go around and talk to people, and then...
then he'd put it all together.
– Harry West, MIT Professor (c. 1990)*

SCIENCE and engineering are reciprocals. Engineering can be seen as a special case of science (“applied science”), but science can equally be described as a special case of engineering (“abstract engineering”). Scientists are given a phenomenon and asked to find its logical and physical relations to the rest of the universe; engineers are given the relations and asked to define the phenomenon. Put differently, scientists derive the specifications from the object, and engineers derive the object from the specifications (Hapgood, F., 1993).

While this dissertation has traits that align with both these definitions of science and engineering, it is clearly biased toward the applied nuance. The main contribution of this research is in the technical progress of PHM in the railway maintenance business. In fact, the “management” pillar of PHM ensures that the technical innovations introduced by the diagnosis and the prognosis pillars have a practical application in the field, otherwise they risk becoming useless and adding no value to address specific problems. No fancy research solution will stand a chance in the market if it is too costly to

implement because it diminishes the return of the investment. This chapter makes explicit the link among the former chapters of this part.

4.1 Literature Review

The state of the art has been divided in two chapters regarding the interest of the reading audience, which is targeted to railway engineers and data scientists.

4.1.1 Railway Engineering

Chapter 2 develops the literature review from the viewpoint of the railway engineer, including general topics such as reliability, management, cybersecurity and sustainability, among others, and also focuses on their application to product development.

The benefits of higher reliability and lower costs can only be realized when the organizational culture adapts to make best use of the technology (Kilian, K., Kilian, M., Mazur, V., and Phelan, J., 2016). In this sense, managing assets and people increasingly gain importance in this context. Reliability and Life Cycle Cost, which lie at the heart of a sustainable railway operation, allow including a cost-benefit analysis into the technical considerations to yield the optimal final maintenance decisions (Mascherona, R., Bellani, L., Compare, M., Trucco, R., Zio, E., 2020). Additionally, with the adoption of information and communications technologies in railway maintenance, vulnerability to cyber threats has increased. Therefore, it is essential that organizations also consider security analytics and automation to increase their resilience against security breaches (Kour, R., Aljumaili, M., Karim, R., and Tretten, P., 2019).

Chapter 2 also determines the main technical problems and open questions in railway engineering, and gives clear indications on the technologies where the maintenance business is going:

- Smart sensors, to collect data from remote locations and/or from distributed machines in continuous movement, such as trains (Boyes, H., Hallaq, B., Cunningham, J., and Watson, T., 2018). Remote monitoring equipment provides a centralized system for fleet diagnostics and supports the shift to predictive maintenance (Smith, K., 2018).
- Machine vision inspection, to concentrate the acquisition of indirect data parameters at the fleet level. The technology provided by machine vision software provides the tools for rail companies to re-

think their approach to reliability engineering (Kilian, K., Kilian, M., Mazur, V., and Phelan, J., 2016).

- Technical language processing, to take advantage of the massive documentation and written work orders on the shop floor. Text-based research and analysis are bound to have a meaningful impact on almost every aspect of the railway sector (Dong, K., Romanov, I., McLellan, C., and Esen, A. F., 2022). They provide an effective approach for improving the collection and analysis of text-based maintenance data, and eventually enable accurate decision-making (Brundage, M. P., Weiss, B. A., and Pellegrino, J., 2020).
- System log analytics, to take advantage of the massive stream of time-stamped event data that the subsystems already generate. The definition of the time intervals in which the failures are likely to occur provides a more intuitive way to handle the predictions by the users, and increases the acceptance of the proposed approaches (Fink, O., Zio, E., and Weidmann, U., 2015c).

4.1.2 Data Science

Chapter 3 adopts the viewpoint of the data scientist, and focuses on the technical topics around Deep Learning, its challenges, and its application on specific approaches to PHM objectives, mainly tackling detection and diagnosis tasks (Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020).

Deep Learning (DL) has gained increasing attention and momentum due to its advantages in data classification and feature extraction problems using neural networks (Schmidhuber, J., 2015; LeCun, Y. and Bengio, Y., and Hinton, G. E., 2015; Raghu, M., and Schmidt, E., 2020). It is an evolving research area with diverse application domains. Hence, it has the potential to increase overall system resilience or cost benefits for maintenance, repair, and overhaul activities (Khan, S., and Yairi, T., 2018).

Chapter 3 also provides indications on the appropriate solutions for product research, emphasizing the usefulness of:

- Multilayer Perceptrons (MLP), as the fundamental building block of neural solutions, exploiting their flexible ability to learn and approximate functions.
- Convolutional Neural Networks (CNN), as the main tool from Deep Learning to deal with signal-like data, and their ability to interpret their operation as filters.

- Autoencoders (AE), as an essential solution for tackling unknown environments through an unsupervised approach.

4.2 Railway Data Operations

Reviewing the potential impacts of the AI technology in the railway maintenance industry (Burroughs, D., 2019c) and business in general (Glover, J., 2013), Table 4.1 charts the operational utilization of railway engineering and data science in the context of railway PHM.

Potential Impact	Railway Engineering	Data Science
Increased capacity	Reliability, Management	MLP
Reduced life cycle costs	Sustainability	MLP
Reduced errors from both humans and existing computer systems	Reliability, Cybersecurity	AE
Improved efficiency and increased performance	Reliability, Management	MLP
High-level automation and autoadaptive systems	Automation, Big Data	MLP
Simplified supervision and fast problem resolution	Automation	CNN
Improved flexibility	Reliability, Management	MLP

Table 4.1 Potential impacts of the AI technology in the railway maintenance industry and business in general.

Table 4.1 shows how the topic of Reliability is intimately linked to Management to address the potential impacts of the increase of capacity and flexibility, regarding the area of Railway Engineering. Similarly, the topic of Automation closely follows to effectively tackle the troubleshooting of errors. Regarding the specific techniques from Data Science, which are not mutually exclusive, the Multilayer Perceptron leads the implementation of solutions for its versatile learning flexibility.

Finally, there is a tight association between the maintenance business activity and the technology that supports the development of the predictive solutions. In this sense, causality works in the two directions: the solutions are affected by the viability of the technology that is available in the market,

and likewise the business is driven by the feasibility of the solutions. Figure 4.1 shows a map of how the assets that lead the development of solutions due to their high return on investment, relate to the specific technologies that are used to develop the added-value predictive maintenance products.

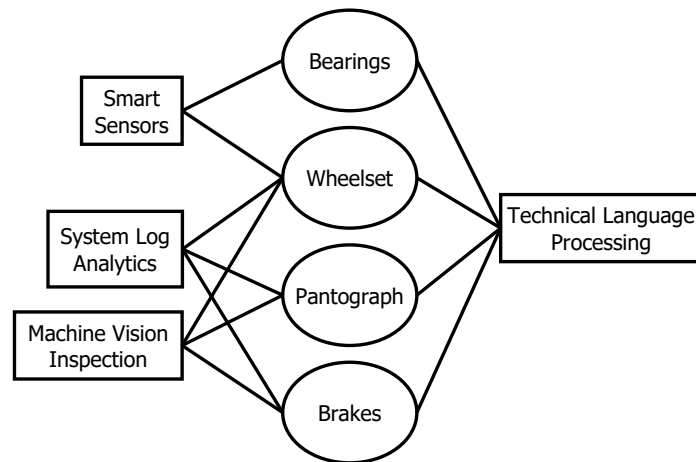


Figure 4.1 Map of railway assets that appeal to business (in circles) and their related predictive maintenance technologies (in squares).

The graphical map of Figure 4.1 shows how specific solutions like the Smart Sensors, which typically operate at the bogie level, i.e., wheelsets and bearings, compare with solutions that feature a broader scope, such as Machine Vision and Log Analytics, which cover the whole energy usage cycle, from its collection (pantograph), exploitation (motor wheelsets for traction), and recovery (regenerative brakes). Finally, the diagram shows how language is regarded to be an adequate medium to tackle the whole train vehicle in a transversal way, potentially taking advantage of wider inter-asset relationships, for example, grasping the extent to which bearings can affect the wheelsets, which in turn may also affect the brakes.

4.3 Applied Research Questions

The generic research question stated in Section 1.2 focuses on building a convincing case for applying DL to the many PHM challenges in the maintenance of railways: Deep Learning displays the characteristics that make it a suitable technology for developing dependable industrial-grade solutions for effectively maintaining rolling stock with confidence.

The next part of the dissertation details the published work introduced in Section 1.4. The salient aspect of the research is in the diversity of the data environments (structured and unstructured) where different DL techniques have been applied with success to solve specific problems and add value to the maintenance business. These distinct scenarios comprise real-valued time-series data (e.g., pantograph carbon strip degradation), real-valued signals (e.g., axlebox mechanical vibration), images (e.g., wheel tread pictures), text (e.g., Return On Experience records), and subsystem events (e.g., blended traction and brake), which have been transformed into time-dependent binary-valued variables. Additionally, the different tasks that have been tackled also support the versatility of DL for PHM. These include the diagnosis of bogies and various subsystems in a broader sense, and the prognosis of pantographs and wheelsets.

What follows is a specific overview of the published work regarding the research question, the state of the art, and how each article contributes to the technical progress of railway PHM.

Enhancing Railway Pantograph Carbon Strip Prognostics with Data Blending through a Time-Delay Neural Network Ensemble

Section 5.1 develops an efficient and robust prognosis solution for the pantograph as a critical railway asset for energy collection, see Section 2.1.10, using a machine vision inspection product to monitor the carbon strip thickness degradation, see Section 2.2.2, and based on a Multilayer Perceptron for its flexibility, see Section 3.1.1. The research question that is investigated can be stated as follows: smartly integrating the unevenly sampled thickness evolution of the carbon strips with external factors that may have an impact on their degradation, such as the seasonal condition of the overhead contact wire, results in more accurate and reliable prognosis. The developed solution can be easily generalized to other friction-driven mechanisms on assets such as the brakes and the wheels, which amount to a big computational load (around 15000 parameter calculations per day at the fleet level).

Pushing Distributed Vibration Analysis to the Edge with a Low-Resolution Companding Autoencoder: Industrial IoT for PHM

Section 5.2 develops a vibration data compression method for the diagnosis of railway axle bearings, see Section 2.1.9, using smart sensors to monitor the mechanical degradation of these bogie components, see Section 2.2.1, and based on a regularized Autoencoder with an undercomplete representation, see Section 3.1.1. The research question that is investigated

can be stated as follows: custom data compression methods are key to enable the remote monitoring and diagnosis of asset condition using devices with a limited data transmission bandwidth. Finally, the learned representation (i.e., the embedding of the Autoencoder) may also be generalized as an encryption method of the data for cybersecurity purposes, see Section 2.1.5.

Integrated Multiple-Defect Detection and Evaluation of Rail Wheel Tread Images using Convolutional Neural Networks

Section 5.3 develops an automatic Deep Learning method to jointly detect and diagnose wheel tread defect images, see Section 2.1.7, using smartphone pictures as a machine vision inspection method, see Section 2.2.2, and based on Convolutional Neural Networks, see Section 3.1.1. The research question that is investigated can be stated as follows: Convolutional Neural Networks lie at the heart of neural solutions and they are able to tackle different tasks and implement the expert troubleshooting process from engineering teams. Since CNN can seamlessly deal with classification and regression objectives from the same filter-interpretable representation, their operation can be easily generalized to tackle any challenge.

Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting

Section 5.4 develops a causal text embedding strategy through a neural encoder, see Section 3.1.1, using a technical language processing solution, see Section 2.2.3, which models the textual entailment of Return On Experience data for bogie diagnosis, see Section 2.1.7. The research question that is investigated can be stated as follows: the useful root cause analysis troubleshooting information is to be found in the entailment of the text, and word embeddings are a feasible solution to extract it. This approach can be used as a building block of more general tasks involving language, such as (causal) reasoning.

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

Section 5.5 develops a method to discover and diagnose anomalies in massive operational data, see Section 2.1.4, using system log analytics, see Section 2.2.4, for subsystem event signals, and through a Variational Autoencoder, see Section 3.1.1. The research question that is investigated can be stated as follows: Variational Autoencoders enable learning many useful hierarchical representations to detect anomalies in data-scarce environments such as the ones typically found in PHM. Additionally, its results yield in-

teresting generalization opportunities for designing Intrusion Detection Systems in the context of cybersecurity, see Section 2.1.5.

PART II

CONTRIBUTIONS

RESEARCH PUBLICATIONS

Any researcher who says that the hottest idea is not what he's working on has got to be stupid. If you think it's the hottest new idea, then why aren't you working on it?
– Geoffrey E. Hinton (1997)

THIS chapter provides the research contributions that have been published during the progress of this doctoral period (2020–2023), see Section 1.4. From the standpoint of the data characteristics, they comprise signal processing (i.e., vibration), time-series of both parametric and nominal variables (i.e., carbon thickness and subsystem events), images (i.e., shop floor pictures), and text (i.e., Return on Experience records). Finally, from the standpoint of applications, the big three objectives of PHM are represented: anomaly detection (i.e., axle box and control network), diagnosis (i.e., wheelset defects and technical-language driven troubleshooting), and prognosis (i.e., pantograph carbon strip wear).

5.1 Conference Paper 1 (2020)

Enhancing Railway Pantograph Carbon Strip Prognostics with Data Blending through a Time-Delay Neural Network Ensemble

This contribution develops a robust prognosis solution for the pantograph based on Multilayer Perceptron, which integrates the thickness of the carbon strips and external factors that may have an impact on their degradation such as the seasonal condition of the overhead contact wire.

This paper was presented on November 2020 at the 12th Annual Conference of the Prognostics and Health Management Society, which was held remotely due to Covid-19 travel restrictions (Trilla, A., Fernández, V., and Cabré, X., 2020).

Enhancing Railway Pantograph Carbon Strip Prognostics with Data Blending through a Time-Delay Neural Network Ensemble

Alexandre Trilla¹, Verónica Fernández², and Xavier Cabré³

^{1,2,3} *Alstom R&D Services, Santa Perpètua de la Mogoda, Barcelona, 08130, Spain*

alexandre.trilla@alstomgroup.com

veronica.fernandez@alstomgroup.com

francesc-xavier.cabre@alstomgroup.com

ABSTRACT

Energy supply for high-speed trains is mainly attained with a high-voltage catenary (i.e., the source on the infrastructure) in contact with a sliding pantograph (i.e., the drain on the rolling-stock vehicle). The friction between these two elements is minimised with a carbon strip that the pantograph equips. In addition to erosion, this carbon strip is also subject to abrasion due to the high current that flows from the catenary to the train. Therefore, it is of utmost importance to keep the degradation of the carbon material under control to guarantee the reliability of the railway service. To attain this goal, this article explores an accurate (i.e., uncertainty bounded) predictive method based on a robust online non-linear multivariate regression technique, considering some factors that may have an impact on the degradation on the carbon strip, such as the seasonal condition of the contact wire, which may develop an especially critical ice build-up in the winter. The proposed approach uses a neural ensemble to integrate all these sources of potential utility with the carbon strip data, which is convoluted in time with a set of spreading filters to increase the overall robustness. Finally, the article evaluates the effectiveness of this prognosis approach with a dataset of pantograph carbon thickness measurements over a year at the fleet level. The results of the analysis prove that it is definitely possible to deploy a fine prediction, and thus yield a new avenue for business improvement through the application of the predictive maintenance approach to pantograph carbon strips.

1. INTRODUCTION

The railway environment in general, and the maintenance of rolling-stock in particular, are recently experiencing great benefits with the deployment of data-driven Prognostics and Health Management (PHM) technology (Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N.,

Alexandre Trilla et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



Figure 1. Alstom TrainScanner deployment at the Manchester Traincare Centre.

2017; Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., and Wang, W., 2015). In line with this source of innovation, Alstom has developed the TrainScanner, which is a track-side train monitoring system that is aimed at optimising the maintenance of brake pads (Trilla, A., Dersin, P., and Cabré, X., 2018), pantograph carbon strips, and wheelsets (Trilla, A., and Cabré, X., 2018), see Figure 1. This product is based on a set of computer vision technologies with lasers and 3D cameras that capture the degradation-related measures for each component as the trains traverse its portal. Then, it automatically triggers the analysis of the collected data, and advises the maintenance team with data-informed prescriptions. This work is particularly focused on the pantograph prognostic enhancement that may be attained with the carbon strip thickness measurements over time.

The British Rail Class 390 rolling stock is an electric high-speed passenger train that conducts the current collection through a pantograph. Therefore, the pantograph is an essential element of the traction chain because it provides access to the power to drive the traction motors, among other

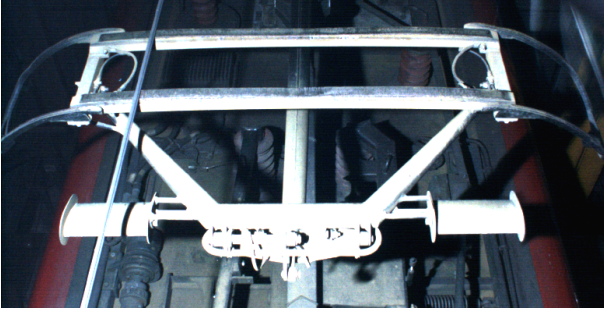


Figure 2. British Rail Class 390 pantograph showing the two carbon strips. One of them leads the contact with the catenary.

systems. In order to draw the current while the train is in motion, the pantograph equips two carbon strips that are in constant sliding contact with the overhead line, also known as the catenary, see Figure 2. Given the permanent friction regime of this means of power transfer, each carbon strip is subject to wear. And in addition to this main degradation mode, there are many other factors that may impact the condition of this asset, such as the amount of current flow (Bucca, G., and Collina, A., 2015; Ding, T., Xuan, W., He, Q., Wu, H., and Xiong, W., 2014), the irregular contact height relative to the rails (Shing, A. W. C., and Wong, P. P. L., 2008), the specific carbon material (Auditeau, G., Bucca, G., Collina, A., and Tanzi, E., 2011; Auditeau, G., 2016), and the ambient temperature (Ocoleanu, C. F., Popa, I., Manolea, G., Dolan, A. I., and Vlase, S., 2009). The combined effect of all these phenomena may produce chips and cracks on the surface of the carbon strip, although the most critical degradation factor that can be directly observed is the season.

This work conducts a thorough analysis of the pantograph carbon strip degradation at the fleet level in order to enhance the performance of its thickness prediction at 30,000 km into the actual operating life of each asset, which is expected to show a great deal of variation according to the seasonal weather. Given the intense mission profile of the fleet, this horizon for the prediction is assumed to provide enough notice time for the maintenance team to schedule the depot resources effectively. The proposed model of the degrading carbon thickness sequence exploits its diversity in time (or distance) through a set of spreading convolutions. Finally, the prognosis evaluation is performed with a rolling window prediction technique, focusing on the uncertainty of the predicted error, which is given by the maximum variability of the error distribution for a given confidence interval.

The article is organised as follows: Section 2 describes the analysis procedure that has been explored, including the description of the data, the evaluation technique, and the prognosis enhancements, along with their preliminary results. Section 3 discusses the overall outcomes and the limitations

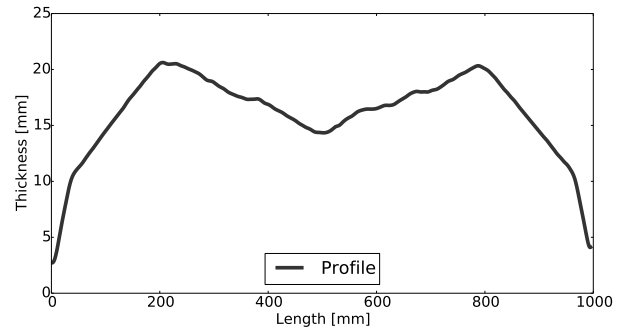


Figure 3. Acquired pantograph carbon strip profile. Note that the figure is not to scale: the carbon strip is a very wide asset.

of the approach, and Section 4 concludes the manuscript and reflects on its impact on the current maintenance plan.

2. METHODS AND RESULTS

This section describes process that has been followed in order to obtain a robust pantograph carbon strip prognosis method. Thus, the development is incremental and preliminary results are provided.

2.1. Carbon Strip Data Preprocessing

The carbon strip is a rectangular piece of carbon material that is mounted at the top of the pantograph. It is 20 mm thick, 30 mm wide and 1,000 mm long. Each pantograph equips two of these strips, and the leader always precedes the contact with the overhead line. Additionally, there are two cars on each train that equip a pantograph, although only one of them is active at a time (i.e., in contact with the catenary). Its rated operating voltage is 25kV AC.

The TrainScanner acquires a cloud of points for each pantograph carbon strip. Based on this data, the carbons are reconstructed with a triangulation technique, and a thickness profile is extracted for each asset, see Figure 3. It can be observed that the degraded area spans from 200 mm to 800 mm, and the most critical part is at the centre, from 400 mm to 600 mm. The system automatically identifies this region and extracts the minimum thickness value for further wear analysis.

This article evaluates the effectiveness of carbon strip prognostics with a dataset of thickness measurements at the fleet level, acquired between June 1 2016 and June 1 2017 at irregular intervals (the monitoring operations are not scheduled). It comprises an amount of 224 strip elements, and each sequence of carbon thickness needs to be preprocessed to add robustness to the prediction. To this end, the following issues are taken into account:

1. **Asset replacement:** steep positive thickness increments (greater than 5 mm) with a final value close to a new asset

measure, i.e., 20 mm, need to be segmented and treated as different assets.

2. **Acquisition failures:** extreme values out of strip range (over 20 mm) or zeroes are regarded as invalid data and thus they need to be discarded from the analysis by removing them from the carbon thickness sequence.
3. **Stability/Monotonicity:** each thickness segment needs to be asserted an overall monotonic negative trend according to the nature of the carbon material erosion, and a monotonic positive progression regarding the accumulated mileage. To this end, a monotonicity index is useful to quantify the amount of regularity in the evolution, which is based on the difference between the number of positive and negative increments (Davydov, Y., and Zitikis, R., 2017).
4. **Sensor precision:** TrainScanner's rated measurement precision is 0.5 mm. The prediction method needs to be robust to this inherent data acquisition system variability.

The resulting set of data should be smooth enough to be subject to further analysis following the ISO 13374 standard (ISO, 2003), which is the main PHM development guideline considered in this work, although similar structured approaches have also been developed for overhead monitoring systems (Brahimi, M., Medjaher, K., Leouatni, M., and Zerhouni, N., 2016). Obviously, the primary interest here is focused on the Prognosis module and the dynamic properties of the carbon strip degradation.

2.2. Rolling Window Prediction Evaluation

A rolling window is a prediction performance estimation procedure that is essentially based on the idea that "the past is used to predict the future". It is an iterative process that frames a history window at some point in the evolution, learns the trend from it in order to make a prediction over a given horizon frame, and finally scores the error difference with the remaining coming data (Hota, H. S., Handa, R., and Shrivastava, A. K., 2007), see Figure 4.

Ultimately, the distribution of the resulting error score is used to estimate the performance of the prediction method, which is mainly driven by the amount of variability (Trilla, A., Dersin, P., and Cabré, X., 2018). To this end, the maximum deviation of the error distribution around its mean value is determined for a confidence interval of 95%. This quantity is here referred to as the "uncertainty". Obviously, the error increases as the prediction horizon is extended into the future.

2.3. Robust Online Linear Regression

The Class 390 tilting Pendolino trains run a steady mission profile on the West Coast Main Line in the UK, featuring a very high availability (running 1,000 miles a day on average), which leads to expect a uniform degradation behaviour. In

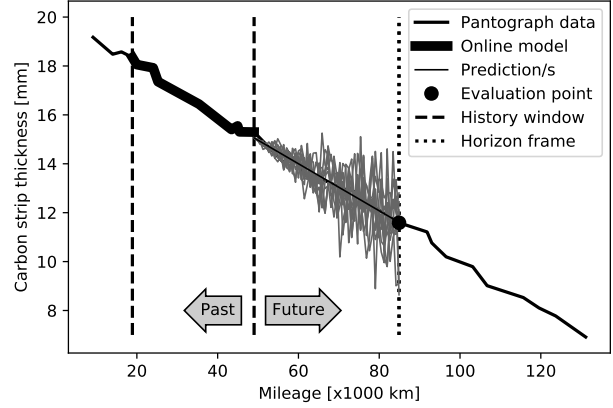


Figure 4. Diagram of the rolling window prediction evaluation for carbon strip data.

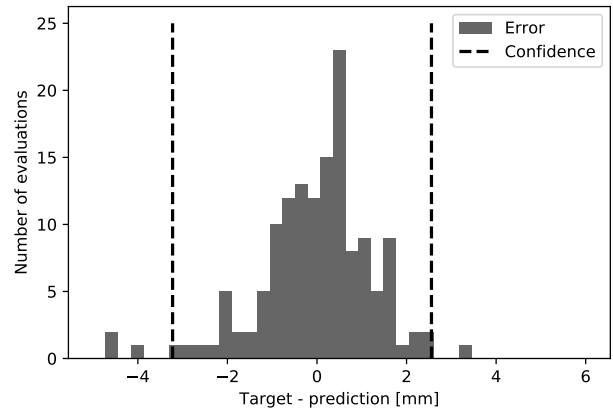


Figure 5. Histogram of the prediction error with robust online linear regression (ROLR). The 95% confidence interval indicates the uncertainty.

order to get a baseline for this study, the model linearity is assumed for the carbon strips in this high-speed rail scenario, following other carbon-based degradations like the brake pads (Trilla, A., Dersin, P., and Cabré, X., 2018). Therefore, a robust ordinary linear regression approach (ROLR) based on weighted least-squares fitting is evaluated. The regression is applied to each window of carbon thickness history after the aforementioned robust data-weighting process, and the prediction is obtained by extrapolating the evolution over the horizon frame. It is to note that the squared-error cost function of use here is very convenient to deal with the data-acquisition precision instability, which may be positive or negative. Finally, given the limited amount of data that is available at the sequence level, the history window is set to be equal to the prediction horizon, i.e., 30,000 km. Figure 5 shows the resulting distribution of this prediction error.

It can be seen that the linear method for the baseline shows an uncertainty of 2.89 mm. However, the resulting distribution shape is asymmetric because it displays a skewed centrality,

instead of the normal Gaussian distribution that would be expected with the least-squares optimisation procedure of use. This might be indicative that the linear assumption is not adequate and perhaps it needs to be questioned. The following sections, though, first delve into the particular bits of information that may be obtained from external context variables, and how they may be used to enhance the prediction.

2.4. Potential Improvement with Seasonal Context

One of the main extrinsic factors that may affect the degradation of the pantograph is the season. Variations of temperature (Ocoleanu, C. F., Popa, I., Manolea, G., Dolan, A. I., and Vlase, S., 2009), humidity, rain, wind... may cause an unsteady wear on the surface of the carbon material of the strip. It is well known that in the winter the contact wire freezes with the icing temperatures, possibly causing abnormal degradation. The spring, instead, is the driest period (although the rain is fairly well distributed throughout the year in the UK).

Further insight into these issues may be displayed through the seasonal wear rates, which grossly indicate the dynamic behaviour of the carbon degradation (i.e., the pace of the deterioration) due to these factors. In order to capture this indicator, the slope parameter of the linear regression on the strip thickness sequence is taken. Figure 6 shows the distribution of wear rates throughout the year using a Gaussian kernel density estimation procedure. It is to note that the winter and spring seasons are located on the extremes of the overall multimodal density. Winter shows the highest rates (over $12 \cdot 10^{-5}$ mm/km), whereas spring shows the lowest rates (under $5 \cdot 10^{-5}$ mm/km). Given that the prediction method of use here is linear (this may be interpreted as the derivative of the wear function), the extremely different error values related to these two sequential seasons prove that a non-linearity is inherently present as seasons gradually change. Therefore, this justifies the specific consideration of the seasonal factor as discrete context variables corresponding to the three modes of wear rate: winter, spring, and summer/autumn (note that their centrality conflates into the same value). The representation of the season as a nominal one-hot encoded vector (instead of a scalar ordinal encoding) is a convenient and effective solution with neural networks (Hancock, J. T., and Khoshgoftaar, T. M., 2020), the use of which is explored further in the following sections.

2.5. Data Blending through Neural Networks

In order to take advantage of the seasonal non-linear context variables discussed in Section 2.4, this section explores blending these different sources of information with a neural network ensemble. Regardless of the difficulty of the prediction task, the neural technique unifies the way of approaching this problem.

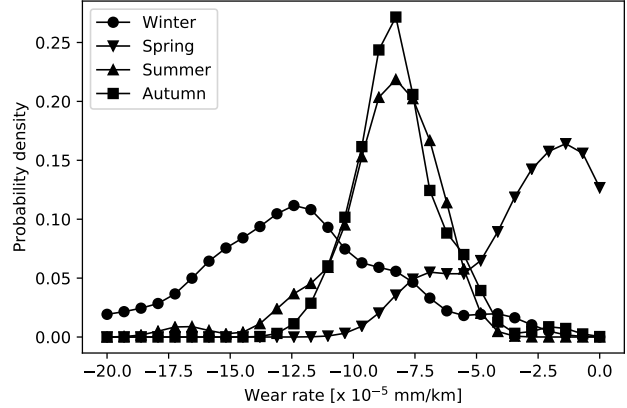


Figure 6. Density distribution of the wear rates according to season throughout the year.

2.5.1. Feature Ensemble with a Multilayer Perceptron

The Multilayer Perceptron is a general-purpose neural network architecture that can seamlessly integrate extrinsic data from different sources in order to refine a prediction (Trilla, A., Dersin, P., and Cabré, X., 2018). It is based on a feed-forward structure with a hidden layer in the middle, which provides the capacity to learn non-linear relationships between the inputs (i.e., the present features) and the output (i.e., the future thickness value). Moreover, its industrialisation is straightforward through a series of matrix multiplications that any platform can efficiently implement with a standard linear algebra library.

For the pantograph carbon strip scenario presented in this work, the baseline prediction result with linear regression is provided as a real-valued feature along with the rest of the aforementioned seasonal context variables (as binary flags with one-hot encoding). Moreover, the strip thickness value within the 30,000 km horizon is provided as the supervised output target prediction, see Figure 7. The hidden neurons are designed with a Rectified Linear Unit activation function to learn the non-linearities (Nair, V., and Hinton, G. E., 2010). The neural network is ultimately trained with a stochastic gradient descent protocol using backpropagation, an adaptive learning rate with momentum (Kingma, D. P., and Ba, J. L., 2015), and considering a squared-error cost function.

In order to get the network to learn effectively, its expressiveness (i.e., the capacity to represent the learnt knowledge) needs to match the complexity of the data within the objective prediction problem. To do so, the number of hidden units H needs to be adjusted because they modulate this learning ability. Note that the input dimensionality of this network is 4 (i.e., 3 context variables plus the result of the linear prediction), therefore, every hidden unit adds 6 new parameters to the model (4 inputs, 1 output, and 1 bias). In order to determine the optimum size of the hidden layer so that underfitting and overfitting learning problems may be

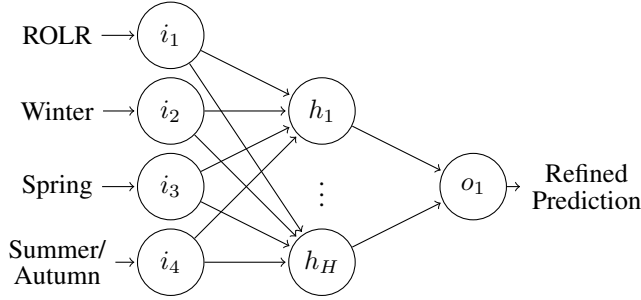


Figure 7. Multilayer Perceptron architecture blending the Robust Online Linear Regression (ROLR) with the set of three seasonal context variables to obtain a better refined prediction.

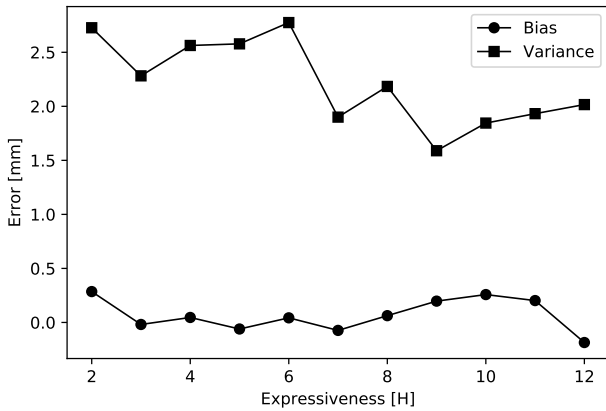


Figure 8. Expressiveness analysis of the neural feature ensemble. The bias represents the mode of the error distribution, and the variance represents its uncertainty.

avoided, a range of values are evaluated with Monte Carlo cross-validation (Dubitzky, W., Granzow, M., and Berrar, D., 2007), applying 10 rounds of repeated random sub-sampling with a train/test split of 95%/5%. This procedure yields over 70 evaluation points, which is a sufficient sample size to reliably estimate the prediction uncertainty. Figure 8 shows the results of this study through a bias/variance tradeoff analysis using the mode and the uncertainty values of the expected skewed error distributions, following customary descriptive statistics tools.

It can be seen that the most interesting performance score (i.e., the variance, or uncertainty) shows a randomly decreasing evolution as the expressiveness of the network grows (i.e., H increases), until the amount of hidden neurons reaches 9. From that point forward, the uncertainty rises, so the network stops generalising and begins to memorise the data, which is a sign of overfitting. Therefore, the optimum size for the hidden layer is of 9 units (it is to note that any residual bias can be corrected a posteriori with this estimation). It can be seen that the resulting system outperforms the previous linear approach as it now shows an uncertainty of 1.59 mm. This improve-

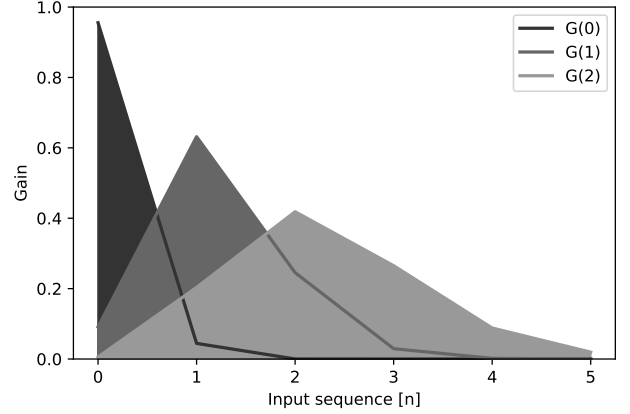


Figure 9. Impulse response of the spreading filters $G(s)$ (with $\alpha = 10$) for the time-delay convolution.

ment is mainly due to modelling the inherent non-linearities in the extrinsic seasonal context variables. Nevertheless, this result is still driven by the assumed linear evolution of the carbon thickness, which is a clear point of improvement that is explored in the next section.

2.5.2. Time-Delay Neural Network Embedding

This section builds upon the former feature ensemble approach, drops the questionable linearity assumption that drives the baseline prediction from Section 2.3, and proposes integrating the carbon thickness data directly through a neural structure known as a Time-Delay Neural Network (TDNN) (Peddinti, V., Povey, D., and Khudanpur, S., 2015). This approach maps the decreasing dynamic evolution of the data into a fixed spatial pattern using a weighted average operation in time with a set of spreading filters $G(s)$ defined by Eq. (1), where L is the size of the delay line (input data buffer), α is the spreading factor, and s is the spatial shift. Note that S is a normalisation factor that ensures that all shifts may deliver the same amount of energy, see Figure 9.

$$G_s = G(s) = \frac{1}{S} \sum_{n=0}^L x[n] \left(\frac{n+1}{s+1} \right)^\alpha \exp \left(-\alpha \frac{n}{s+1} \right) \quad (1)$$

In addition to empowering the system to deal with the thickness data evolution directly (i.e., an autoassociation that does not assume any specific behaviour, like the linearity), the convolution with the spreading filters exploits the local features of the data and reduces the searchable weight space for the learning stage. Furthermore, it increases the robustness to uneven sampling, which is to be taken into account as the inspections through the TrainScanner are not scheduled. This, in turn, enables the neural network that follows to handle sequences with different lengths, which is a clear limitation of

the ordinary multilayer perceptron (where the input dimensionality is fixed). Also, the use of variable history lengths may be of help to reduce the high uncertainty of strips showing a faster wear rate (Trilla, A., Dersin, P., and Cabré, X., 2018; Greitzer, F. L., and Ferryman, T. A., 2001).

The enhanced solution that this work suggests first builds the time-series embedding by applying the filters over the thickness sequence to obtain three spatial shifts (i.e., the high, middle, and low parts of the evolution). It uses the spreading factor α as a modulator to adjust the bandwidth of the filters to the length of any given sequence (applying the first filter $G(0)$ to the newest thickness sample to deal with a most unweighted value close to the prediction result), see Figure 9. And then, it assembles the resulting physical features with the former set of seasonal context variables that have proven to be useful in this modelling approach. Figure 10 shows this architecture.

At this point, the expressiveness of the new multilayer perceptron needs to be adjusted to the new embedded features following the cross-validation procedure described in Section 2.5.1. Now, each hidden unit adds 8 new parameters to the model. Figure 11 shows the result of this expressiveness analysis, which indicates that with 6 hidden neurons, the uncertainty of the prediction drops to 1.39 mm. Note that for this richer input representation (6 variables instead of 4), the model has become somewhat simpler (6 hidden units instead of 9), which makes perfect sense regarding the complexity tradeoff between the features and the predictive learning capacity.

3. DISCUSSION

This work exposes the gradual performance enhancement of pantograph carbon strip prognosis, initially relying on linear regression (resulting in 2.89 mm of uncertainty), then refining this prediction by accounting for non-linearities through the seasonal context information (1.59 mm), and finally dealing with the thickness evolution data directly with a set of spreading filters (1.39 mm). What is more, if these error results are assumed to belong to a normally distributed random variable, their incremental differences are statistically significant with a confidence interval of 95% using an Independent Samples t-test. In this case, the powerful Student hypothesis test with the Gaussian normality assumption is preferred over weaker non-parametric approaches like the Mann-Whitney U test, in spite of its apparent appropriateness to compare skewed distributions.

Despite the nice interpretability of the initial linear behaviour that emulates the prominent uniform physical degradation of this asset, every step taken toward dropping this linear assumption has led to increasingly better results in terms of prediction uncertainty. However, the resulting neural model has also increased its complexity, thus becoming more difficult to

interpret. Neural networks are typically regarded as “black boxes” because of their intricate nested inner functions.

In order to shed some light into the internal behaviour of the best-performing TDNN ensemble model, Figure 12 shows an input-standardised sensitivity analysis based on the profile method (Shojaeefard, M. H., Akbari, M. Tahani, M., and Farhani, F., 2013). It can be seen that there are three correlated patterns of behaviour:

- The three physical features (low, middle, and high-parts of the thickness sequence) show a rather linear increasing pattern along 8 mm of the whole output dynamic range. Their likelihood can be explained by their common source of information (i.e., the carbon evolution), which is already expected to be linearly uniform.
- The winter and spring seasons show a convex function (first negative, and then positive), with the inflection point around 0.5σ , and also covering 8 mm of the output dynamic range. These two seasons display the most extreme wear rates, see Figure 6, and the neural network seems to use them in a similar way for the refined predictions. In the end, it's in the transition from winter to spring that the main nonlinearity occurs.
- The summer/autumn seasonal variable exhibits a kind of offset rectifier function with the inflection point located at -0.5σ . This variable stretches up to 12 mm of the output dynamic range. It is to note that the associated wear rate applies to six months and it is represented by one single variable, thus maybe this explains its extended range.

While it may be difficult to assess the contribution of each variable in terms of importance, the amount of dynamic range in the output may be indicative of their rank, leaving the summer/autumn flag as the most critical variable. Further testing with an ablation study would be needed to derive stronger statements.

The current approach conducts a rough discretisation of the seasonal factor with three mutually-exclusive binary variables, but seasons change gradually, and the mid-season nuances are possibly missed with this solution. Nevertheless, conducting a seasonal information blending, e.g., at the month level, increases the number of extrinsic variables from three to twelve, and this in turn may enlarge the amount of weights in the neural network to an excess of expressiveness, increasing the potential risk of overfitting the data.

In addition to the principal seasonal information, other external sources of potential prognosable input have also been informally studied. On the one hand, there is the particular location of the pantograph. Each Class 390 train equips two pantographs, and the decision of using one or the other depends on the driver. This arbitrary factor may affect the degra-

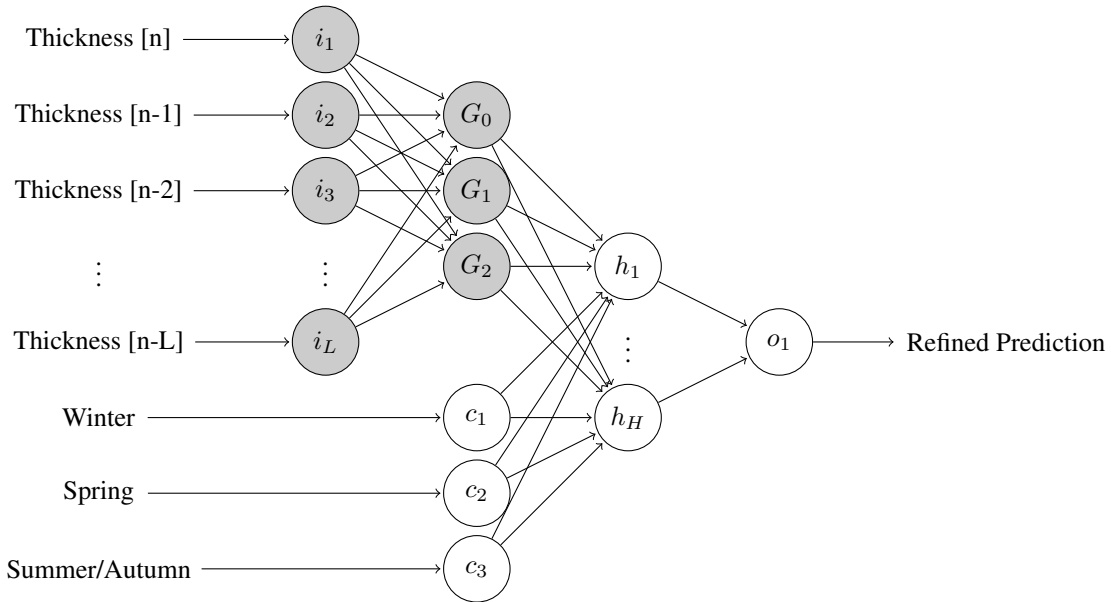


Figure 10. Architecture of the Time-Delay Neural Network ensemble. Strip thickness data is convoluted with the spreading filters (shown as shaded units), and blended into the set of context variables with a multilayer perceptron.

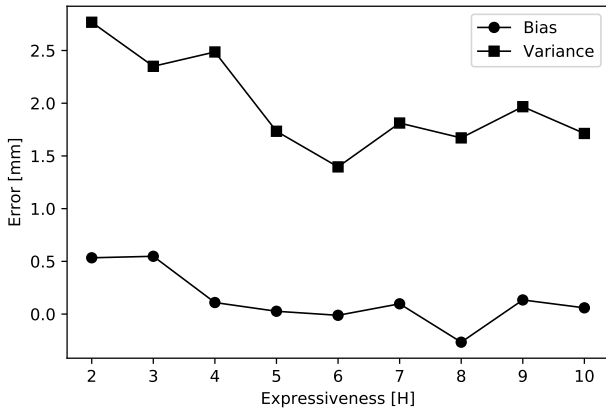


Figure 11. Expressiveness analysis of the TDNN ensemble. The bias represents the mode of the error distribution, and the variance represents its uncertainty.

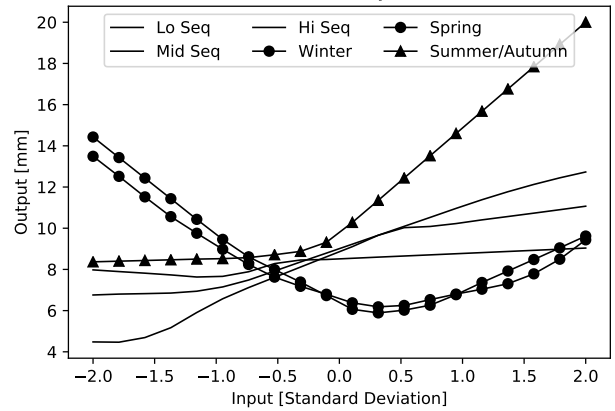


Figure 12. Sensitivity analysis of the best-performing TDNN ensemble ($H=6$). The dynamic range of the inputs is normalised to their standard deviation. The variables that display the same pattern share the same line style.

dation of the carbon strips, although particular behaviours seem unlikely to be displayed because driver rota is the common way of operating the rolling stock.

On the other hand, there is the position of the carbon strip in the pantograph. Depending on the sense of the trip (upwards to Scotland, or downwards to England), different strips lead the contact with the catenary. But again, it's the driver's decision to use one pantograph or the other, so for the same rotation reason, a singular behaviour is unlikely to show. In

the end, neither the pantograph location nor the strip position have proven to be of much use in the prognostication of future carbon strip thickness.

4. CONCLUSION

At present, the carbon strip replacement criterion for the Class 390 pantographs is based on a single thickness threshold value. This inefficient approach does not take into account

the rate of wear that the different strips display, which varies significantly throughout the year with the seasons. Thus, the same thickness value can lead to different operating mileages before the asset reaches its actual end of life (i.e., when there is no carbon material left on the strip).

This article presents the most sophisticated technique for TrainScanner pantograph carbon strip prognostics, which is based on a Time-Delay Neural Network that blends a spread sequence of carbon thickness values with the seasonal context information. This approach yields a prediction error uncertainty around 1.39 mm at the asset level and for a projected horizon of 30,000 km, which is related to the planning time that is necessary for scheduling the maintenance resources at the depot. Therefore, if the expected mileage to the next visit is under this distance frame, the strip threshold scrap limit could be safely extended up to this performance value.

The future work that is currently envisaged may further deal with other extrinsic context variables in order to add more robustness to the prognosis method. The neural network has proven to be a very versatile approach for assembling different data sources. In this regard, we may exploit the temporal persistence of large amounts of other nominal (i.e., non-parametric) data provided by related onboard subsystems (Hu, X., Eklund, N., and Goebel, K., 2007), e.g., from traction. Alternatively, we also expect to explore other sequence learning approaches through the Long Short-Term Memory units (Hochreiter, S., and Schmidhuber, J., 1997), and seek the complementary characteristics that may help the current approach attain a better effectiveness.

ACKNOWLEDGMENT

We would like to show our gratitude to Professor Xavier Vilasís-Cardona for his valuable support and comments that greatly improved the manuscript.

REFERENCES

- Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N. (2017). Prognostics and Health Management for Maintenance Practitioners - Review, Implementation and Tools Evaluation. *International Journal of Prognostics and Health Management*, 8(60), 1–31.
- Auditeau, G. (2016). Effect on wear of contact strips and wire of materials used for pantograph sliding strips (COSTRIM project). *Proc. of the World Congress on Railway Research*.
- Auditeau, G., Bucca, G., Collina, A., and Tanzi, E. (2011). Experimental analysis of effect of plain carbon and impregnated carbon contact strips on contact wire wear. *Proc. of the World Congress on Railway Research*.
- Brahimi, M., Medjaher, K., Leouatni, M., and Zerhouni, N. (2016). Critical Components Selection for a Prognostics and Health Management System Design: an Application to an Overhead Contact System. *Annual Conference of the Prognostics and Health Management Society (ISBN: 978-1-936263-22-6)*, 7, 1–8.
- Bucca, G., and Collina, A. (2015). Electromechanical interaction between carbon-based pantograph strip and copper contact wire: A heuristic wear model. *Tribology International (ISSN: 0301-679X)*, 92, 47–56.
- Davydov, Y., and Zitikis, R. (2017). Quantifying non-monotonicity of functions and the lack of positivity in signed measures. *Modern Stochastics: Theory and Applications*, 4(3), 219–231.
- Ding, T., Xuan, W., He, Q., Wu, H., and Xiong, W. (2014). Study on Friction and Wear Properties of Pantograph Strip/Copper Contact Wire for High-Speed Train. *The Open Mechanical Engineering Journal (ISSN: 1874-155X)*, 8, 125–128.
- Dubitzky, W., Granzow, M., and Berrar, D. (2007). Fundamentals of data mining in genomics and proteomics. *Springer Science & Business Media*, 178.
- Greitzer, F. L., and Ferryman, T. A. (2001). Predicting Remaining Life of Mechanical Systems. In *Intelligent Ship Symposium IV*.
- Hancock, J. T., and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 7(28), 1–41.
- Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(1), 1735–1780.
- Hota, H. S., Handa, R., and Shrivastava, A. K. (2007). Time Series Data Prediction Using Sliding Window Based RBF Neural Network. *International Journal of Computational Intelligence Research*, 13(5), 1145–1156.
- 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.
- ISO. (2003). *Condition monitoring and diagnostics of machine systems: Data processing, communication and presentation* (Tech. Rep. No. 13374-1:2003). International Organization for Standardization.
- Kingma, D. P., and Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs.LG]*, 1–15.
- Nair, V., and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proc. of the 27th International Conference on Machine Learning*, 1–8.
- Ocoleanu, C. F., Popa, I., Manolea, G., Dolan, A. I., and Vlase, S. (2009). Temperature investigation in contact pantograph AC contact line. *International Journal of Circuits, Systems and Signal Processing (ISSN: 1998-4464)*, 3(3), 154–163.
- Peddinti, V., Povey, D., and Khudanpur, S. (2015). A Time Delay Neural Network Architecture for Efficient Mod-

eling of Long Temporal Contexts. In *Interspeech* (pp. 3214–3218).

- Shing, A. W. C., and Wong, P. P. L. (2008). Wear of pantograph collector strips. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit (ISSN: 2041-3017)*, 222(2), 169–176.
- Shojaeefard, M. H., Akbari, M. Tahani, M., and Farhani, F. (2013). Sensitivity Analysis of the Artificial Neural Network Outputs in Friction Stir Lap Joining of Aluminum to Brass. *Advances in Materials Science and Engineering*, 1–7.
- Trilla, A., and Cabré, X. (2018). Determining the Equivalent Conicity for Railway Wheelset Maintenance with Deep Ensembles. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 10(1), 1–6.
- Trilla, A., Dersin, P., and Cabré, X. (2018). Estimating the Uncertainty of Brake Pad Prognostics for High-speed Rail with a Neural Network Feature Ensemble. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 10(1), 1–7.
- Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., and Wang, W. (2015). Prognostics and Health Management: A Review on Data Driven Approaches. *Mathematical Problems in Engineering*, 2015(793161), 1–18.

BIOGRAPHIES

Alexandre Trilla graduated from La Salle University of Barcelona with a M.Sc. in Electronics and Telecommunications 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 (*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 the solutions with artificial neural networks.

Verónica Fernández graduated from Universitat Autònoma of Barcelona with a B.Sc. in Chemical Engineering in 2016, and a M.Sc. in Automatic Systems and Industrial Electronics Engineering from Universitat Politècnica of Catalonia in 2019. At present, she is a Data Scientist working on PHM in the railway environment.

Xavier Cabré was born in Barcelona. He graduated from La Salle University of Barcelona with a M.Sc. in Electronics and Telecommunications Engineering in 1998. He has an industrial research background in high-technology environments. In 2015, he co-authored a patent in image processing and computer vision for the detection of railway components. At present he is the Project/Program manager of TrainScanner.

5.2 Conference Paper 2 (2020)

Pushing Distributed Vibration Analysis to the Edge with a Low-Resolution Companding Autoencoder: Industrial IoT for PHM

This contribution develops a vibration data compression method for the diagnosis of railway axle bearings using a regularized Autoencoder with an undercomplete representation. Additionally, the embedding of the Autoencoder may be regarded as an encrypted representation of the data for cybersecurity purposes.

This paper was presented on November 2020 at the 12th Annual Conference of the Prognostics and Health Management Society, which was held remotely due to Covid-19 travel restrictions (Trilla, A., Miralles, D., and Fernández, V., 2020).

Pushing Distributed Vibration Analysis to the Edge with a Low-Resolution Companding Autoencoder: Industrial IoT for PHM

Alexandre Trilla¹, David Miralles², and Verónica Fernández³

^{1,3} *Alstom, Santa Perpètua de la Mogoda, Barcelona, 08130, Spain*
alexandre.trilla@alstomgroup.com
veronica.fernandez@alstomgroup.com

² *Grup de Recerca en Tecnologies Media, La Salle - Universitat Ramon Llull, Barcelona 08022, Catalonia, Spain*
david.miralles@salle.url.edu

ABSTRACT

The Industrial Internet-of-Things (IIoT) has disrupted the way of collecting physical data for predictive maintenance purposes. At present, networks of intelligent wireless sensors are pervasive, finding success in many environments and industries, including the railways. However, when it comes to data-intensive applications like vibration monitoring that require the delivery of large amounts of records, the limitations of these devices arise. The shortfalls are mainly driven by the low-bandwidth transmission capacity of their radio interfaces, and the low-power features of their battery-operated (and/or energy-harvested) electronics. In sight of these limited resources, this article explores a vibration data compression strategy for diagnosis purposes. To maximise the amount of transferred information with the least amount of bytes this method works in three stages: first, it extracts the most useful features for vibration-based analytics. Then, it compresses the raw signal waveform using an Autoencoder neural network with an undercomplete representation, assessing its optimum regularisation approach: the denoising, sparse, and contractive configurations. Finally, it reduces the resolution of the compressed data by quantising all the resulting real values into single-byte unsigned integers. The proposed strategy is evaluated with a dataset of railway axle bearings with different levels of degradation. The results of the analysis show that with compression rates up to 10 the vibration signals are practically unaffected by this procedure, and once the signals are reconstructed with a minimum quality standard, many diagnosis goals like anomaly detection, fault location, and severity appraisal can be performed. This approach yields a wide range of business opportunities for on-board predictive maintenance with IIoT technology.

Alexandre Trilla et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

The Industrial Internet-of-Things (IIoT) has emerged as one of the leading technologies to deploy the remote condition monitoring of machines (Boyes, H., Hallaq, B., Cunningham, J., and Watson, T., 2018), especially when such machines are transportation assets that move around the territory. This work is particularly concerned with the application of Prognostics and Health Management (PHM) to the maintenance of mechanical rolling-stock components (Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N., 2017), specifically those able to be inspected with vibration-monitoring technology. In this regard, Alstom has developed The Motes (Trilla, A., and Gratacòs, P., 2016, 2013), which is a network of intelligent wireless sensors that capture the vibration signature of such mechanical elements and provide feedback about their actual degradation stage, see Figure 1. These sensors have been designed to acquire vibration in different operational regimes, both on the workshop floor (low-speed environment) and in commercial service (up to high-speed rail). Ultimately, the fleet management team can take advantage of their added-value and make better informed decisions on how to schedule the various maintenance actions with the available resources. In this setting, one of their main objective components are the axle bearings, also known as axleboxes.

The axlebox is a heavy-duty safety-critical railway element (Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., and Wang, W., 2015). It bears the weight of the train, minimises the friction with the rotating axle, and its failure in service might cause derailment. Therefore, its maintenance is of utmost importance to guarantee the availability of the fleet. To this end, in a predictive maintenance scenario, the collected vibration signature must be reliable and truly representative of the actual degradation of the asset. However, this often comes at the cost of transmitting a greater amount of data, i.e., its raw

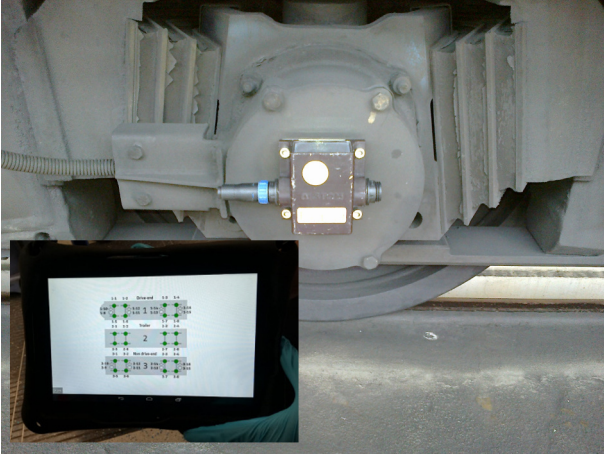


Figure 1. The Motes in use with an axlebox at the Morden depot in London for the Northern Line underground. The small window at the bottom-left corner also shows a tablet, which is used to manage the network of sensors.

signal waveform. Relaying big loads of data works against the business-case for the IIoT, especially for remote battery-powered devices, which are designed with wide-range but low-bandwidth and low-energy radio interfaces, and are expected to operate intermittently to last a long time unattended. In addition, the activity of the sensors must not delay the limited time of the maintenance staff during their inspection actions. Overall, this exposes the need to maximise the throughput of information with the smallest volume of vibration data, and to do so, this article explores the use of signal compression as a key enabler to achieve a cost-effective, robust, and easy-to-implement PHM solution (Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., and Wang, W., 2015).

In the context of wireless sensor networks for diagnosing machinery, vibration signal compression has already been attained using different signal processing methods like the Discrete Cosine Transform (Alsalaet J. K., Najem, S. I., and Ali, A. A., 2012), the Empirical Mode Decomposition (Chan, J. C., and Tse, P. W., 2009), and Wavelets (Hao, W., and Jinji, G., 2012). However, the electronics used for some IIoT devices populate low-power processors that aim at the minimisation of energy consumption at the expense of featuring somewhat modest processing capabilities. Thus, implementing such costly complicated time-frequency transforms is oftentimes out of reach. In this regard, this article proposes the use of neural networks as a general-purpose function approximator because of their overall good effectiveness, and also because their industrialisation reduces to making use of linear algebra operations like matrix multiplication and vector addition, which are already widely supported by many embedded platforms. Specifically, the proposed approach focuses on using the Autoencoder neural network.

The Autoencoder is a particular layered neural architecture

that inherently learns to replicate data through a compressed representation. Its previous use in PHM highlights its capacity to detect anomalies (Goldthorpe, P., and Desmet, A., 2018) and to construct health indices (Trilla, A., Janjua, F., and Bermejo, S., 2019), among others. This article uses the compressed layer of the Autoencoder to obtain a condensed description of the raw signal waveform, which is the most critical factor in terms of transmitted data volume. Additionally, a set of vibration health features are also extracted and appended to the compressed signal to refine its eventual expanded reconstruction. The computational cost of this stage is not relevant in this context, but the amount of computed indicators must be kept to a minimum to reduce the amount of transmitted data. Finally, this array of information is quantised into a low-resolution single-byte representation to build a compact frame for the IIoT infrastructure, thus attaining the goal of transmitting a high-quality vibration signal with a fraction of the originally acquired data sample.

The article is organised as follows: Section 2 describes the distributed compression/expansion analysis procedure, including the Autoencoder technique, and the description of the railway axlebox data. Section 3 shows the results of the signal reconstruction evaluation. Section 4 discusses the overall approach, and Section 5 concludes the manuscript, reflects on its impact to the current maintenance actions, and provides avenues of future improvement.

2. METHOD

This section describes the process that has been followed to obtain a reliable vibration compression procedure.

2.1. Distributed Vibration Compadding

In order to reduce the amount of transmitted data while retaining the fundamental characteristics of the vibration signal, the whole process needs to be split into the following functions:

- **Compression** of the time-varying signal waveform and its features on the edge (i.e., the sensing device).
- **Expansion** of the compressed signal and its feature-corrected reconstruction on the user side (i.e., the cloud, or a mobile platform like a tablet).

Figure 2 shows the complete compadding procedure (note that “compadding” is the portmanteau of “compressing” and “expanding”). The specific operations performed by the edge device for the compression stage are described as follows:

1. **Data Acquisition** The sensing device equips an accelerometer that is used to obtain an instance of the vibration signature for the degrading asset (e.g., the axlebox). The dynamic range of the sensor and the sampling frequency of use are adjusted to the test conditions (i.e., at the depot or in commercial service). A sequence of

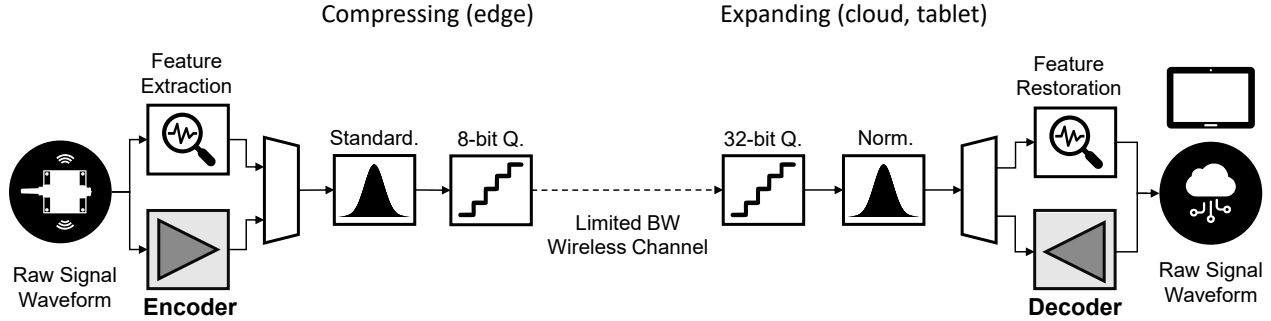


Figure 2. Diagram of distributed vibration compression and expansion processes for transmitting information over a limited bandwidth (BW) wireless channel. The role of the encoder and the decoder (both implemented with the Autoencoder) operating on the raw signal waveform (most critical data volume) is highlighted in boldface.

real-valued samples are collected; thus, a signed 32-bit floating-point arithmetic is used.

2. **Feature Extraction** An array of statistical health indicators for vibration data are extracted, e.g., peak magnitude, variance, skewness, kurtosis, crest factor, etc. (Trilla, A., Janjua, F., and Bermejo, S., 2019; Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., and Wang, W., 2015). These features describe particular aspects of the asset’s degradation (e.g, driven by the failure modes).
3. **Encoding** The stream of raw vibration waveform data is segmented into short-time windows, and each of these frames is then compressed with the Autoencoder, yielding a fraction of the initial acquisition size. The next section provides further details about this operation.
4. **Standardisation** Each of the variables obtained so far (the features and the compressed vibration map) is statistically standardised so that their resulting distribution has zero mean and unit standard deviation (a Gaussian shape is also assumed), i.e., $\mathcal{N}(0, 1)$. This process is also known as Z-score normalisation.
5. **8-bit Quantisation** The resulting real values are finally rescaled so that the ultimate normal distribution is centred on the 0-255 value range. Therefore, each variable now has a $\mathcal{N}(128, 64^2)$ distribution, which is discretised and may be represented with an unsigned 8-bit integer arithmetic after a rounding operation, thus obtaining a low-resolution representation. It is to note that this final step requires the truncation of the standardised distribution to fit into the limited range of the single byte representation. The truncated range is arbitrarily set to cover 95% of the real values (i.e., 2 standard deviations).

Similarly, the specific operations performed by the end user device for the expansion stage (i.e., the cloud or a mobile platform like a tablet) reverse the process described above: first, the low-resolution samples are quantised into a real-valued 32-bit floating-point arithmetic. Then, the original variable distributions are normalised, which recovers the vi-

bration features directly. And finally, the encoded waveform values are decoded into the initial vibration signals with the Autoencoder. It is to note that this is a lossy compression procedure, so one last post-processing step is applied to ensure that the reconstruction preserves the original health features. In this work, the peak magnitude of the vibration is maintained because it is mostly indicative of the severity of the incipient failure.

2.2. Autoencoder Neural Network

The Autoencoder (AE) is a connectionist machine learning technique that replicates “essential information”. It is data-specific, so it only works with instances that are of same nature as the examples it has learnt from. To this end, it uses a self-supervised learning technique that exploits auto-association (Kramer, M. A., 1992; Stone, V. M., 2008), which is a specific mode of supervised learning where the targets are generated from the inputs. As a result, this neural network learns a distributed representation of the data that captures its meaningful attributes as its main factors of variation (Bengio, Y., 2009).

For the end-to-end vibration compression purposes that this work pursues (implemented on the edge device, and on the cloud/tablet), the design of the proposed neural network architecture is feed-forward and shallow, i.e., memoryless with one single hidden layer. This reduces both the memory footprint and the computational burden, and the resulting weights that define the behaviour of the model may be directly industrialised through a set of matrix multiplications (Goldthorpe, P., and Desmet, A., 2018). In addition, the framework of the presented Autoencoder shows a converging layout from its input dimensionality D into H at half of its depth (i.e., the encoding, compression stage), and then a diverging structure back to D toward its output (i.e., the decoding, expansion stage), see Figure 3. This undercomplete configuration forces the Autoencoder to learn the most salient features of the training data, and thus it develops a compressed repre-

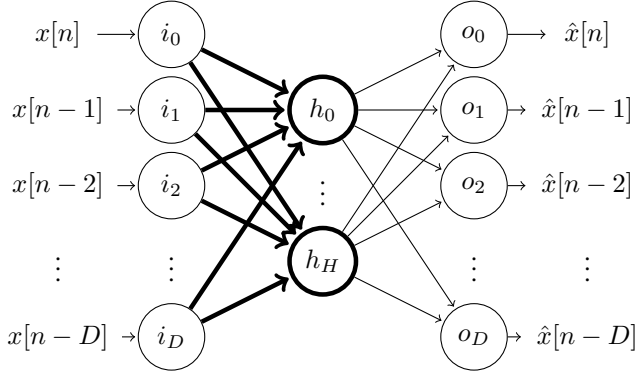


Figure 3. Compadding Autoencoder architecture. D is the input data dimensionality, and H is the size of the hidden/encoding layer, which defines the learning capacity of the neural network. The undercomplete representation is ensured as long as $H < D$. The encoder part is shown with thick arrows (along with thick states for the compressed vector), whereas the decoder is shown with thin arrows.

sensation. The amount of hidden units H in the “bottleneck” layer, which must be smaller than D in this case, defines the expressiveness of this neural network and therefore modulates its learning capacity. Additionally, if these hidden neurons apply a nonlinear activation function like a Rectified Linear Unit, the network gains the ability to capture multi-modal aspects of the input distribution (Japkowicz, N., Hanson, S. J., and Gluck, M. A., 2000), although in this case the compression transformation is essentially linear (from input to hidden layer). Obviously this Autoencoder-based approach is lossy, in the sense that the replica only retains the principal characteristics of the data, but not the details (or the noise). A greater reconstruction quality may be obtained with the identity, the principal component, or the overcomplete representations (making H equal to or greater than D), but these would clearly work against the compression objective.

The proposed Autoencoder is trained with an advanced stochastic gradient descent procedure with backpropagation following the Adam algorithm (Kingma, D. P., and Ba, J. L., 2015), which implements the weight updates through the individual estimation of the first and second statistical moments of the gradients (i.e., a momentum on the gradient and its squared value). The specific hyperparameters of use are: a learning rate α of 0.001, a first momentum β_1 of 0.9, and a second momentum β_2 of 0.999. The average root mean square (RMS) error between the reconstruction and the original vibration signal is used as the objective cost function, i.e., $(\hat{x}[n] - x[n])^2$. This conventional optimisation protocol still has room for some improvements through regularisation penalties, yielding different Autoencoder solutions. These refinements are described hereunder.

2.2.1. Ordinary AE

This Autoencoder is directly trained to compress the input into some lower-dimensional representation so that the *exact same input* may thereafter be reconstructed, without any further constraint. This is analogous to a maximum-likelihood estimation of the optimum weights, and therefore it is subject to overfitting. Obviously, some kind of regularisation strategy would be desirable here, but the Ordinary AE does not contemplate it explicitly; this model only relies on the limited representational capacity of the undercomplete architecture. However, this work also exploits the advantage of limiting the number of epochs during training, because gradient descent with early stopping is similar to a squared Euclidean norm regularisation of the weight parameters (Zinkevich, M., 2003). This strategy generalises the performance of the resulting Autoencoder.

2.2.2. Denoising AE

Another strategy for regularising the Autoencoder is by stochastically corrupting the vibration signal input with noise, while the original uncorrupted signal is still used as target for the reconstruction. This method is known as the Denoising AE (Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A., 2010). This approach learns to preserve the statistical regularities of the input vibration signal, and to undo the random corruption, which can take different forms:

- **Additive White Gaussian Noise (AWGN)** The addition of wideband noise is inspired by many natural processes, and its Gaussian amplitude distribution is driven by the central limit theorem of probability theory when many random processes interact. This is a basic noise model used in information theory, and this work regards its useful convenience for the corruption of the input.
- **Masking** The random setting of some inputs to zero is also a successful regularisation method. This occlusion strategy forces the Autoencoder to deal with data that contains missing values. This is an interesting property because it regards the Autoencoder as a generative model.

2.2.3. Sparse AE

Another strategy for regularising the Autoencoder is via the sparsity in the encoding space. The Sparse AE (Makhzani, A., and Frey, B., 2014) offers an alternative method for constraining the amount of information that may traverse the network and thus require a learned compression of the input data, without reducing the number of hidden units. This Autoencoder adds a sparsity penalty on the activation of the hidden layer so that only a few units may operate at a given time (the correction is increased with the amount of contribution). In this approach, the network gets selective and sensitive to individual hidden units toward specific attributes of the input

vibration data. This sparsity cost is attained by computing the average activations in the hidden layer, and then scoring the Kullback-Leibler divergence between a Bernoulli random variable with this mean value, and another one with a desired small sparse average value.

2.2.4. Contractive AE

There is yet another strategy for regularising the Autoencoder considered in this work that is known as the Contractive AE (Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y., 2011). In this approach the Autoencoder is trained so that the derivatives of the hidden layer activations are small with respect to the input. This prevents that small changes in the input may lead to large changes in the encoding space, so in a sense it adds robustness to small perturbations around the data. This effect is attained by introducing a penalty term in the cost function that corresponds to the Frobenius norm of the Jacobian matrix of the encoder activations with respect to the input. It is shown that this results in a localised space contraction, which in turn yields robust features on the activation layer.

2.3. Vibration Data and Stream Processing

In the present PHM environment, real-time data exchange is not necessary because the gradual degradation of mechanical assets like axleboxes does not occur in a short time. Thus, The Motes operate with asynchronous connectivity (Boyes, H., Hallaq, B., Cunningham, J., and Watson, T., 2018). However, the compression feature of the Autoencoder is limited to its input dimensionality D . In order to transmit a whole “long” vibration signal as a stream, the original sequence needs to be buffered and segmented into windows of length D , then compressed into vectors of length H , and finally be transmitted sequentially in the payload of the wireless protocol frames for the available interfaces, e.g., Wi-Fi, ZigBee, Bluetooth LE, or LoRa.

To evaluate the effectiveness of the companding method with the Autoencoder, this work uses a dataset of axlebox vibration data acquired for a metro stock, rolling at 5mph, on a straight level test track, in the depot. Each acquisition comprises a waveform of 4 seconds sampled at 3200Hz. The complete dataset includes over 28000 instances of vibration segments (with 500 samples each) divided into different degradation levels (Trilla, A., Janjua, F., and Bermejo, S., 2019), i.e., good, regular, and bad condition.

3. RESULTS

This section compares the different Autoencoder strategies to determine which of them yields the best companding effectiveness for the IIoT, i.e., the maximum compression with the minimum loss. Their performance is estimated with a round of stratified random subsampling with 5% of the instances

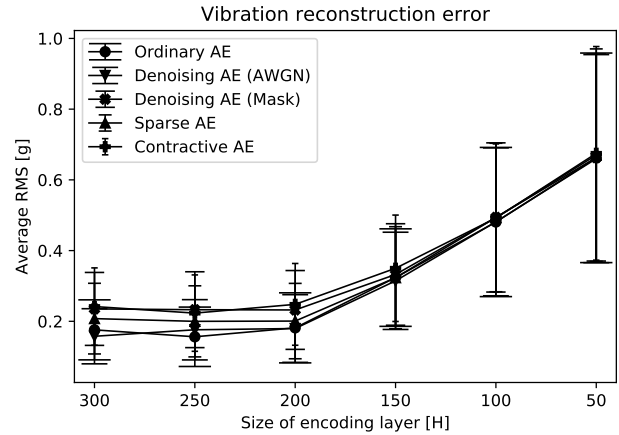


Figure 4. Autoencoder reconstruction error with respect to the size of the hidden/encoding layer (H). The points correspond to the mean value of the RMS error distribution (assuming Gaussian normality), and the whiskers correspond to one standard deviation. Note that the AE strategy of use may be distinguished by the shape of the points and the size of the error caps.

(i.e., around 1400) for testing. Figure 4 shows how the size of the hidden/encoding layer impacts the reconstruction error of the test signals for each AE approach.

In general, it can be seen that regardless of the regularisation strategy of use, all approaches display a flat constant error response down to 200 hidden units (with a greater or lesser offset), and a linear increasing slope beyond that inflection point (also increasing the variability). The interpretation that follows for this effect is that down to 200 hidden units the Autoencoder generalises well, but further compression limits its representational capacity to a point that the neural network underfits the data and so exhibits a steady increase of the reconstruction error. Additionally, it is the Ordinary Autoencoder, which only relies on the undercomplete representation for regularising its performance, the one that attains the lowest reconstruction error. When an additional regularisation strategy is applied, the resulting “over-regularised” Autoencoder diminishes its ability to adapt and converge to a better solution. Taking the inflection point at $H=200$ hidden units as the reference (with input $D=500$), the difference between the least performing strategy (i.e., the Contractive AE, with $\mathcal{N}(0.2479, 0.1156^2)$) and the best (i.e., the Ordinary AE, with $\mathcal{N}(0.1815, 0.0991^2)$) is statistically significant with a confidence interval of 95% using an Independent Samples t-test.

It is to note that this reconstruction performance is averaged over all test instances, which belong to different condition categories. In order to shed some light into this particular aspect, Figure 5 shows the distribution of error values regarding the degradation of the test assets for the best-performing companding strategy, i.e., the Ordinary Autoencoder with 200

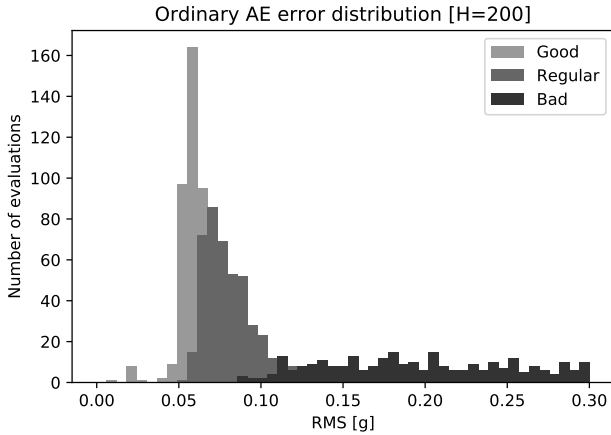


Figure 5. Ordinary AE (with input $D=500$, and compressed output $H=200$) reconstruction error with respect to degraded test asset conditions (good, regular, and bad).

hidden units. This graph makes it clear that as the axleboxes degrade, the reconstruction accuracy of the Autoencoder decreases, and that happens precisely for the most critical situations, when warnings and alarms possibly need to be raised (i.e., for the bad condition). That's why it is of utmost importance to take into account the health features to refine the reconstruction of the waveforms. This loss of reconstruction performance with the progress of the degradation is probably caused by the increased dynamic range and non-stationarity of the signals. In addition, the shape of this distribution questions the previous normality assumption, so the former results must only be taken as indications.

Finally, the transformation of a window of 500 vibration samples into a condensed vector of 200 points yields a compression rate of 2.5, and the 8-bit quantisation that follows applies another rate of 4. Therefore, the final compression rate is of 10, and the resulting system displays a good (almost lossless) companding performance. Figure 6 and Figure 7 show how the Ordinary Autoencoder reconstructs a vibration signal in the worst-case scenario: foreshadowing a failure (the original signal belongs to the "bad" axlebox condition). It can be seen that the time waveform preserves the amplitude that signals the severity of the degradation, and the frequency spectrum retains the location of the source of the failure, so the signal compression process does not modify the result of the analysis that would be obtained with the original raw data. In the healthy case, where the discrepancy between the original waveform and its reconstruction is even smaller, a complete overlap is visually observed, with a signal amplitude an order of magnitude smaller. Consequently, the Ordinary Autoencoder approach enables a fine-grained diagnosis through IIoT monitoring technology.

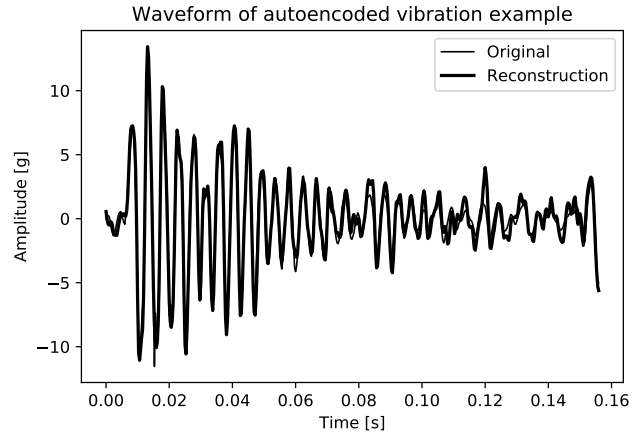


Figure 6. High-peaked non-stationary time waveform of an autoencoded vibration signature showing a bad condition (Ordinary AE with $D=500$ and $H=200$).

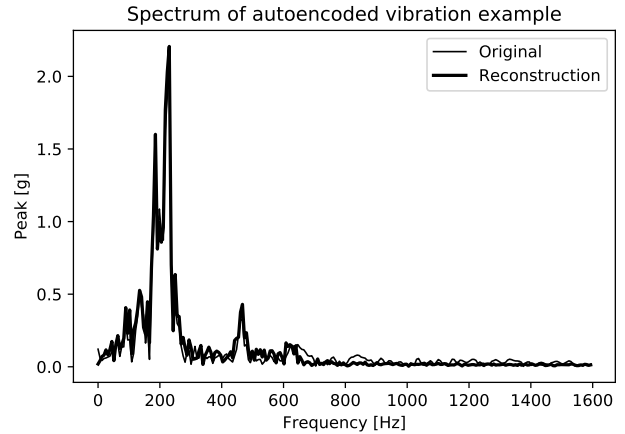


Figure 7. Frequency spectrum of the autoencoded vibration signature shown in Figure 6.

4. DISCUSSION

By trying to approximate the identity function with an under-complete representation, the Autoencoder attains a flexible compression strategy that significantly reduces the amount of data to be transmitted. However, the Autoencoder is not usually considered to be a good compressor in the conventional broad sense, because it lacks the versatility to be applied to data of arbitrary nature. It doesn't operate by exploiting the redundancy in the data to build efficient codewords, so perhaps its performance is limited by this aspect. Nevertheless, it is to note that the compressed layer of the Autoencoders studied in this work corresponds to the linear components of the vibration signals (Duda, R. O., Hart, P. E., and Stork, D. G., 2001), and on that space a clustering technique followed by vector quantisation could still be applied to obtain such an encoded codebook of principal centroids (despite possi-

bly preventing the detection of anomalies in this latent representation). Additionally, the low-resolution quantisation step presented in this work is linear, and a more effective procedure might be obtained with a nonlinear quantiser enhancing the main concentration of data in the feature distribution.

In order to better understand the internal behaviour of the Autoencoder beyond the mapping, other strategies have also been considered, like the use of convolutions and filters. Inspired by the suggestion that the architecture of the neural network is more important than the values of the weights (Gaier, A., and Ha, D., 2019), the use of pairwise correlations has been studied to exploit sparse time dilations like WaveNet (Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K., 2016) and Time-Delay Neural Networks (Peddinti, V., Povey, D., and Khudanpur, S., 2015). In the case that the vibration waveform gets averaged as if by the use of a low-pass filter, the fundamental signal behaviour is retained, but the Autoencoder increases the reconstruction error with an offset. Similarly, the same result is obtained if the input waveform is down-sampled to enhance the details contained in the high-frequency components. In both cases, though, the performance inflection point at 200 hidden units is equally obtained. Therefore, it seems that the densely layered Autoencoder eventually learns the most effective signal transformation, but as the compression rate is incremented, the reconstruction is increasingly smoothed (Trilla, A., Janjua, F., and Bermejo, S., 2019).

Finally, it is to note that the current compression is obtained with a linear combination of the input vibration samples, which is similar to the data-driven measurement matrix that may be developed in compressed sensing (Wu, S., Dimakis, A. G., Sanghavi, S., Yu, F. X., Holtmann-Rice, D., Storcheus, D., Rostamizadeh, A., and Kumar, S., 2019). The recent state of the art applied to vibration signals (which also involves frequency considerations) obtains compression rates up to 5 (Premanand, B., and Sheeba, V. S., 2020), whereas the approach described in this contribution reaches rates of 10 with the same error. However, the inclusion of an additional hidden layer before (and after) the current encoding would lead to an intricate nonlinear representation, potentially smaller than 200 units, and therefore increase the current compression rate. The universal approximation theorem suggests that this is possible (Cybenko, G., 1989), but it has not been explored in this work to minimise the processing especially on the edge device. In a similar vein, the space complexity is also to be considered in an embedded Machine Learning environment given the limited memory of some microcontrollers (Warden, P., and Situnayake, D., 2020). The largeness of the encoding matrix, thus, may be a limiting factor of the industrial deployment of this solution. Nonetheless, this size may be conveniently reduced by shortening the length of the input buffer while keeping the same compression

rate at the expense of increasing the running time, e.g., compressing 250 vibration samples into 100 (instead of 500 into 200) maintains the same representational capacity with a quarter of the original matrix size (in number of weights), and it takes twice as much to complete the processing.

5. CONCLUSIONS

The use of the activation in the hidden/encoding layer of an Ordinary Autoencoder with an undercomplete representation along with a low-resolution quantisation step, significantly reduces the amount of vibration data to be transmitted through an IIoT monitoring network. With compression rates up to 10, the high quality of the reconstructed signal waveforms permits implementing a fine-grained diagnosis. The proposed approach reduces the needed bandwidth for the transmission, and/or shortens the download time for each acquisition. Also, its impact speeds up the maintenance cycle on the workshop floor, and/or increases the inspection frequency on remote locations.

The future work that is currently envisaged opens up two main fronts. On the one hand, exploring the use of complex numbers to obtain a richer representational capacity of the underlying neural network (Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J. F., Mehri, S., Rostamzadeh, N., Bengio, Y., and Pal, C. J., 2018). And on the other hand, developing a deep network pruning strategy to facilitate its implementation on embedded systems with limited hardware resources (Han, S., Mao, H., and Dally, W. J., 2016).

ACKNOWLEDGMENT

We would like to show our gratitude to Professor Xavier Vilasís-Cardona for his valuable support and comments that greatly improved the manuscript.

REFERENCES

- Alsalaet J. K., Najem, S. I., and Ali, A. A. (2012). Vibration data compression in wireless sensors network. *Proc. of the International Conference on Signal Processing, Communication and Computing*, 717–722.
- Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N. (2017). Prognostics and Health Management for Maintenance Practitioners - Review, Implementation and Tools Evaluation. *International Journal of Prognostics and Health Management*, 8(60), 1–31.
- Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2(1), 1–71.
- Boyes, H., Hallaq, B., Cunningham, J., and Watson, T. (2018). The industrial internet of things (IIoT): An

- analysis framework. *Computers in Industry*, 101, 1–12.
- Chan, J. C., and Tse, P. W. (2009). A Novel, Fast, Reliable Data Transmission Algorithm for Wireless Machine Health Monitoring. *IEEE Transactions On Reliability*, 58(2), 295–304.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2, 303–314.
- Duda, R. O., Hart, P. E., and Stork, D. G. (Ed.). (2001). *Pattern Classification*. Wiley-Interscience.
- Gaier, A., and Ha, D. (2019). Weight Agnostic Neural Networks. *Proc. of the 33rd Conference on Neural Information Processing Systems*, 1–19.
- Goldthorpe, P., and Desmet, A. (2018). Denoising autoencoder anomaly detection for correlated data. *Proc. of the Fourth European Conference of the Prognostics and Health Management Society*, 1–6.
- Han, S., Mao, H., and Dally, W. J. (2016). Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *Proc. of the International Conference on Learning Representations*, 1–14.
- Hao, W., and Jinji, G. (2012). The research of optimal selection method for wavelet packet basis in compressing the vibration signal of a rolling bearing in fans and pumps. *Proc. of the 25th International Congress on Condition Monitoring and Diagnostic Engineering, Journal of Physics: Conference Series*, 364(012033), 1–13.
- Japkowicz, N., Hanson, S. J., and Gluck, M. A. (2000). Non-linear autoassociation is not equivalent to PCA. *Neural Computation*, 12(3), 531–545.
- Kingma, D. P., and Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs.LG]*, 1–15.
- Kramer, M. A. (1992). Autoassociative Neural Networks. *Computers and Chemical Engineering*, 16(4), 313–328.
- Makhzani, A., and Frey, B. (2014). k-Sparse Autoencoders. *Proc. of the International Conference on Learning Representations*, 1–9.
- Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs.SD]*, 1–15.
- Peddinti, V., Povey, D., and Khudanpur, S. (2015). A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts. In *Interspeech* (pp. 3214–3218).
- Premanand, B., and Sheeba, V. S. (2020). Compressed encoding of vibration signals using extremum sampling. *Springer Nature Applied Sciences*, 2(1261), 1–10.
- Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011). Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. *Proc. of the 28th International Conference on Machine Learning*, 833–840.
- Stone, V. M. (2008). The auto-associative neural network - a network architecture worth considering. *Proc. of the 2008 World Automation Congress*, 1–4.
- Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J. F., Mehri, S., Rostamzadeh, N., Bengio, Y., and Pal, C. J. (2018). Deep Complex Networks. *Proc. of the International Conference on Learning Representations*, 1–19.
- Trilla, A., and Gratacòs, P. (2013). Condition based maintenance on board. *Chemical Engineering Transactions Journal*(33), 733–738.
- Trilla, A., and Gratacòs, P. (2016). Maintenance of bogie components through vibration inspection with intelligent wireless sensors: a case-study of axle-boxes and wheel-sets using the Empirical Mode Decomposition technique. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*(230), 1408–1414.
- Trilla, A., Janjua, F., and Bermejo, S. (2019). Developing a Hybrid Expert/Data-Driven Health Index for Railway Axleboxes Using Auto-encoder Neural Networks. *Proc. of the Prognostics and System Health Management Conference*, 1–6.
- Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., and Wang, W. (2015). Prognostics and Health Management: A Review on Data Driven Approaches. *Mathematical Problems in Engineering*, 2015(793161), 1–18.
- 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.
- Warden, P., and Situnayake, D. (2020). *TinyML: Machine Learning with Tensorflow Lite on Arduino and Ultra-Low-Power Microcontrollers*. UK: O’Reilly UK Ltd.
- Wu, S., Dimakis, A. G., Sanghavi, S., Yu, F. X., Holtmann-Rice, D., Storcheus, D., Rostamzadeh, A., and Kumar, S. (2019). Learning a Compressed Sensing Measurement Matrix via Gradient Unrolling. *Proc. of the 36th International Conference on Machine Learning*, 1–17.
- Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. *Proc. of the 20th International Conference on Machine Learning*, 928–936.

BIOGRAPHIES

Alexandre Trilla graduated from La Salle University of Barcelona with a M.Sc. in Electronics and Telecommunications Engineering in 2008, and a M.Sc. in IT Manage-

ment 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 (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 the solutions with artificial neural networks.

Dr. David Miralles holds a degree on Theoretical Physics of the University of Barcelona (1995). From 1996 to 2001 he worked at the Fundamental Physics Department at the same university. In 2001 he obtained a PhD on Mathematical Physics. From 2001 to 2007 he worked at the Depart-

ment of Communication and Signal Theory of Ramon Llull University (URL). He has made several stays in international centres: Instituto de Matemática, Estatística e Computação Científica, Campinas, (Brazil, 1998); International Center of Theoretical Physics, Trieste (Italia, September 2004); Observatoire de Paris (France, March 2005-06) and MIT Media Lab, Cambridge (USA, May 2014). He is member of Grup de Recerca en Tecnologies Media where he leads the Interaction area; a research team focused on designing new interactions between people and objects.

Verónica Fernández graduated from Universitat Autònoma of Barcelona with a B.Sc. in Chemical Engineering in 2016, and a M.Sc. in Automatic Systems and Industrial Electronics Engineering from Universitat Politècnica of Catalonia in 2019. At present, she is a Data Scientist working on PHM in the railway environment.

5.3 Journal Article 1 (2021)

Integrated Multiple-Defect Detection and Evaluation of Rail Wheel Tread Images using Convolutional Neural Networks (2021)

This contribution develops an automatic Deep Learning method to jointly detect and diagnose wheel tread defect pictures using Convolutional Neural Networks.

This article was published on May 2021 in the International Journal of Prognostics and Health Management (Trilla, A., Bob-Manuel, J., Lamoureaux, B., and Vilasis-Cardona, X., 2021).

Integrated Multiple-Defect Detection and Evaluation of Rail Wheel Tread Images using Convolutional Neural Networks

Alexandre Trilla^{1,4}, John Bob-Manuel², Benjamin Lamoureux³, and Xavier Vilasis-Cardona⁴

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

² *Alstom, Morden, London, SM45PT, United Kingdom*
john.bob-manuel@alstomgroup.com

³ *Alstom, Saint Ouen, Paris, 93482, France*
benjamin.lamoureux@alstomgroup.com

⁴ *DS4DS, La Salle, Universitat Ramon Llull, Barcelona, 08022, Spain*
xavier.vilasis@salle.url.edu

ABSTRACT

The wheel-rail interface is regarded as the most important factor for the dynamic behavior of a railway vehicle, affecting the safety of the service, the passenger comfort, and the life of the wheelset asset. The degradation of the wheels in contact with the rail is visibly manifest on their treads in the form of defects such as indentations, flats, cavities, etc. To guarantee a reliable rail service and maximize the availability of the rolling-stock assets, these defects need to be constantly and periodically monitored as their severity evolves. This inspection task is usually conducted manually at the fleet level and therefore it takes a lot of human resources. In order to add value to this maintenance activity, this article presents an automatic Deep Learning method to jointly detect and classify wheel tread defects based on smartphone pictures taken by the maintenance team. The architecture of this approach is based on a framework of Convolutional Neural Networks, which is applied to the different tasks of the diagnosis process including the location of the defect area within the image, the prediction of the defect size, and the identification of defect type. With this information determined, the maintenance-criteria rules can ultimately be applied to obtain the actionable results. The presented neural approach has been evaluated with a set of wheel defect pictures collected over the course of nearly two years, concluding that it can reliably automate the condition diagnosis of half of the current workload and thus reduce the lead time to take maintenance action, significantly reducing

engineering hours for verification and validation. Overall, this creates a platform of significant progress in automated predictive maintenance of rolling stock wheelsets.

1. INTRODUCTION

Wheel tread degradation is a common downtime cause for rolling-stock which can significantly affect service availability. Railway wheelsets are usually made of steel because of the high load they must bear and the generally high speed of this transport service. In this setting, it is in the wheel-rail interface that the incipient degradation damage develops as visible defects like cracks, spalls, shells, and skid flats (Magel, E., and Kalousek, J., 1996). If the severity of these defects compromises the safety operational considerations of the railway service (among other additional criteria, like the comfort of the passenger in high-speed rail), the trains are driven out of commercial service to perform a reprofiling maintenance action with the lathe in the depot. This activity is typically scheduled on a periodic mileage basis, but due to the nature of defect occurrence and its evolution, inspections are carried out as part of the regular maintenance procedure to guarantee the reliability of the service and extend the wheel life.

The inspections of wheel tread condition have been traditionally approached by monitoring dynamic variables (i.e., time-varying signals) such as the force and the strain, and also by using static variables like the raster image provided by a picture, which is rich in spatial information. And regarding the data-driven algorithms of their diagnosis methods, most of these strategies rely either on low-level/pixel-wise heuristics (Zhang, W., Zhang, Y., Li, J., Gao, X., and Wang, L., 2014;

Alexandre Trilla et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
<https://doi.org/10.36001/IJPHM.2021.v12i1.2906>

Hyde, P., Ulianov, C., and Defossez, F., 2016; Zhang, J., Guo, Z., Jiao, T., and Wang, M., 2018), or on maximum-margin classifiers like the Support Vector Machines (SVM) (Ma, K., Vicente, T. F. Y., Samaras, D., Petrucci, M., and Magnus, D. L., 2016; Guo, G., Peng, J., Yang, K., Xie, L., and Song, W., 2017). However, these solutions seem to be complementary, and neither clearly outstands its counterpart.

Out of the numerous endeavors to detect rail wheel defects, this work underlines the study developed by Krummenacher and colleagues, which compares an approach using wavelets with SVM to a time-series embedding with a Convolutional Neural Network (CNN), motivated by the recent success of this widely-adopted deep neural Computer Vision technology (Krummenacher, G., Ong, C. S., Koller, S., Kobayashi, S., and Buhmann, M., 2018). Their investigation concludes that the CNN approach improves the classification performance through its automatic representation learning ability. This result is much in line with the current popular Machine Learning (and in particular Deep Learning) research trend driven by CNN's ability to spot surface degradation problems (Han, K., Sun, M., Zhou, X., Zhang, G., Dang, H., and Liu, Z., 2017; Shang, L., Yang, Q., Wang, J., Li, S., and Lei, W., 2018; Zhang, Y., Cui, X., Liu, Y., and Yu, B., 2018).

Following state of the art Deep Learning techniques for Prognostics and Health Management (PHM) (Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020), the present work is concerned with the design and implementation of a rail wheel tread defect diagnosis system based on CNN applied to smartphone pictures that is able to cope with the increasing productivity demand to maintain more assets with the same resources and reduce the engineering lead time to take maintenance action (Vickerstaff, A., Bevan, A., and Boyacioglu, P., 2020). To attain this goal, this approach breaks down the complexity of the whole value chain into modules that may be developed in their own specific context, and it blends the hands-on experience available on the shop floor with the strong technical background available in the engineering office. In addition, an industrialized online web application based on modern software development tools and practices is also created to deploy this solution at the fleet level.

This article outlines the different steps involved in the development of this project: from the research that statistically states the feasibility of the proposed solution, to its industrialization through a minimum-viable product as a proof of concept. Section 2 describes the design procedure, including the description of the data, the learning technique and its evaluation, and the robust industrialized solution. Section 3 shows the expected performance results. Section 4 discusses the overall outcomes and the limitations of the approach, and Section 5 provides the conclusions of the work and reflects on its impact on the current maintenance plan along with the future avenues of improvement.

2. METHOD

This section describes process that has been followed to obtain a robust wheel tread defect diagnosis method.

2.1. Defect Data Description

To merge the knowledge from both the depot workshop and the engineering office, data from each environment needs to be available for learning. This section describes the kind of information that can be extracted from each perspective.

2.1.1. Maintenance Data

A collection of 4600 wheel tread defect pictures taken with smartphones has been compiled over the course of two years by the maintenance repair and overhaul (MRO) team in the Alstom's Traincare Centre (i.e., the London Underground Northern Line fleet). The maintenance staff take pictures whenever an incipient defect is detected on the shop floor. The accumulated dataset depicts the presence of six different defects, which are described as follows with increasing severity:

Indentation (INDT) Superficial dent caused by the wheels running over a hard object on the track. This category also includes the "pitting" defect, which displays a similar effect on the wheel tread but its root cause is the mechanical strain.

Rolling Contact Fatigue (RCF) Cracks caused by repeated contact stress during the rolling motion of the wheels. RCF is a major wear issue in the London Underground infrastructure and its monitoring is incredibly labor intensive requiring precise visual inspection and detailed data recording (Vickerstaff, A., Bevan, A., and Boyacioglu, P., 2020).

Wheel Flat (FLT) Rash that appears on both wheels caused by the wheelset skidding on the rail.

Clustering (CLUS) Also known as multiple cavities, it has to do with the appearance of several bruises along the tread due to uneven contact issues.

Spalling (SPALL) Also known as single cavity or shelling, it is the critical development of one of the multiple cavities described before.

Crazing (CRAZ) Also known as thermal cracking, it is a fracture that occurs with repeated heating and cooling of the wheel tread surface caused by traction and braking actions.

As an example, Figure 1 shows a wheel tread picture with spall and RCF defects. In this dataset, though, there is a strong bias toward the RCF type (with over 80% of the instances). Such a major defect type imbalance may pose an adverse situation for Machine Learning (Yang, Y., and Xu, Z., 2020). Therefore, this work downsamples the RCF subset of data so that the resulting defect type distribution is more amenable to

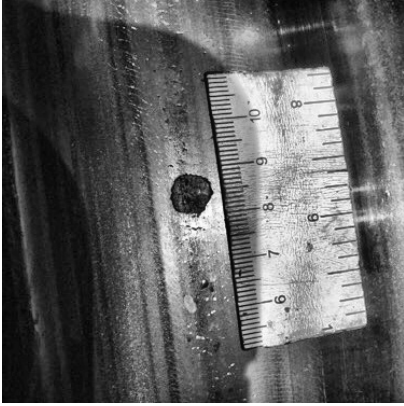


Figure 1. Wheel tread showing SPALL and RCF defects.

Table 1. Wheel tread defect dataset properties, including the number of instances, the defect type distribution, the contextual information (i.e., location and physical size of the defect), and the diagnosis assessment.

Attribute	MRO Data	ENG Data
Size	1200	118
INDT	23%	16%
RCF	36%	52%
FLT	21%	7%
CLUS	8%	11%
SPALL	11%	4%
CRAZ	1%	0%
(None)	0%	10%
Location	✓	✗
Physical size	✓	✗
Go	✗	15%
Warning	✗	51%
Stop	✗	34%

direct supervised learning. The reduced working maintenance dataset comprises 1200 picture instances, and its new defect type distribution is shown in Table 1. It can be seen that the crazing defect type is the underrepresented minority with only 1% of the instances. This skew is likely to cause some learning trouble, but that's an inherent difficulty in this environment that the proposed system will evaluate.

In addition to the graphical content of the picture, the maintenance staff also provides additional information in the form of textual data, identifying the inspected train unit, the physical size of the defect, etc. This unstructured context is processed with regular expressions to deal with the uneven spacing, the letter casing, etc., in order to complement the description of the spotted defects. Nevertheless, the dependability in this supplementary material may be questionable, and the picture remains to be the most reliable datum that the engineering team reviews for the definitive diagnostic. Therefore, the MRO context must only be used as an informative indication.

2.1.2. Engineering Data

In a similar vein, the engineering (ENG) team has curated a collection of 118 defects, see Table 1 for details. Note that this dataset is an order of magnitude smaller, and also exhibits a strong bias toward the RCF defect type. In addition, this set misses the “crazing” type, and it contains the absence of defect (i.e., images without a problem).

Although contextual data such as the location and the physical size of the defect are not available here, what is especially important is the condition assessment from the expertise, which also displays strong bias toward the “warning” statement. This is the engineering advice that drives the maintenance actions. In sight of the characteristics of the MRO and ENG datasets, which are both partially overlapping and complementary, there may exist some potential criteria transfer issues that need to be observed.

2.2. Image Processing

The collection of raster images that depict the wheel tread defects poses challenging issues due to the variability of the hand-held smartphone-based capture process. Depending on who is taking the picture and when, there is inconsistency in the focus, distance to the defect, lighting, etc. To address these concerns, a pixel-level Image Processing module is created.

2.2.1. Preprocessing

First, the three color channels (i.e., RGB) are conflated into one single intensity channel. The steel of the wheel treads is mostly blue-grayish, and any decoloration in the metal is equally visible with a shade on the resulting black-and-white picture, so the useful information is expected to be retained with this transformation. The image is now computationally lighter and therefore more tractable for further analysis.

Then, the edges of the picture, which may be taken vertically or horizontally, are trimmed so that the resulting image is standardized with a squared shape. Note that the area of interest containing the defect is always located around the center. With this operation, the size of the image is reduced to three quarters of its original size, which adds yet another time-computational advantage as less data needs be processed.

Finally, the histogram of the image is equalized to enhance its contrast (Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B. t. H., Zimmerman, J. B., and Zuiderveld, K., 1987). Figure 1 illustrates the application of these preprocessing steps to a defective wheel tread picture.

2.2.2. Data Augmentation

The abundance of data is required to design a Computer Vision solution based on Deep Learning, and the current defect

data collections are insufficient for use according to modern dataset size standards. In this situation, the system is likely to overfit and memorize the data, thus lacking the capacity to generalize. Therefore, a series of affine transformations (i.e., modifications that preserve the collinearity and the ratios of distances) are applied to these instances in order to augment their amount while retaining the salient degradation information (Simard, P. Y., Steinkraus, D., and Platt, J. C., 2003). Specifically, 4 translations (north, south, east, and west shifts), 2 rotations (clockwise and counterclockwise), and 4 mirrorings (horizontal, vertical, and the combined flipping) are performed. Additionally, 2 levels of additive white Gaussian noise are also applied. Eventually, the size of the dataset is increased 64-fold, yielding a working collection of over 80k instances (original and manufactured), which now enables exploring the data-driven solution. What is more, it is known that even small input perturbations like these are sufficient to considerably degrade the system's performance (Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A., 2019). Therefore, by taking them into account during the training procedure, the final system is expected to increase its overall robustness against these potentially adverse effects (Hermann, K. L., Chen, T., and Kornblith, S., 2020).

2.3. Multitask System Architecture

To tackle the complexity of the wheel tread defect diagnosis problem, this work suggests a divide-and-conquer approach, where the main task is divided into five specialized data-driven modules:

Defect Detection - Location (DD-Loc) Identifies the central point of the defect area in the preprocessed image. This task is addressed as regression problem (i.e., landmark detection) where the coordinates of the defect location are predicted in pixel space.

Defect Detection - Physical size (DD-Phy) Predicts the size of the defect (width and length) in a given physical measure (e.g., millimeters). This task is also addressed as a regression problem.

Defect Classification (DC) Discriminates the different types of defects present in the defect area of the input picture. This task is addressed as a multi-label classification problem where the defects are not mutually exclusive, and the outputs represent defect membership probabilities. Ultimately, these probabilities are rated against a threshold θ_{DC} and a discrete vector of potential defects is issued.

Engineering Assessment (EA) Determines the diagnostic based on the type of defect, its physical size, and a set of embedded logical rules that guarantee the minimum acceptance criteria. The output complies with a kind of traffic lights interface: go, warning, and stop.

Confidence Index (CI) Indicates the degree of trustworthiness in the provided diagnosis. Its output operates as a

binary variable.

Figure 2 shows the end to end diagnosis chain. Note that in addition to these five main data-driven modules, there is also the Image Processing (IP) block (already explained in Section 2.2), the defect cropping block, and the circumference calculation (CC) block. The latter two auxiliary blocks are self-explanatory.

2.4. Convolutional Neural Networks

The task division approach ensures that the multiple sources of learning signals do not get scrambled, so that each module can specialize. However, all these detection and classification problems operating on image data can be solved effectively with a convolutional neural architecture, mimicking the hierarchical feature learning strategy that occurs with the visual system's compositional structure (Bengio, Y., 2009), where the initial layers learn basic forms and the subsequent layers combine them to create complex patterns. CNN's are exceptionally successful at dealing with the high dimensionality of an image because they inherently reduce the solution search space (i.e., amount of learnable parameters) with a weight sharing strategy: they use a series of trainable filters that exploit the local surface statistical regularities of the pictures (Jo, J., and Bengio, Y., 2017), making the whole neural system less prone to overfit the data. In turn, this approach also makes these networks especially robust to location, detecting the same pattern in different parts of the photographs as the same filter kernel is reused throughout the image, which exhibits a translationally invariant structure.

Given all this common framework, this section describes a single flexible unified CNN to be applied to each task independently. For computational purposes there is an implicit image rescaling to 75 px that does not compromise the details of the defects, as the spatial aggregation of lower dimensional embeddings can be done without much or any loss in representational power (Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., 2016).

2.4.1. Framework Layout

Discovering neural network architectures remains a laborious but crucial task (Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y. L., Tan, J., Le, Q. V., and Kurakin, A., 2017), because carefully balancing network depth, width, and resolution can lead to better performance (Tan, M., and Quoc, V. L., 2019). In the aim of taking advantage of the many years of focused investigation in neural layouts, the proposed CNN framework is fundamentally based on the classic LeNet-5 architecture (Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P., 1998), which defines two convolutional stages and three fully connected stages, and the AlexNet architecture (Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012), which includes some Deep Learning improvements like the Rectified Lin-

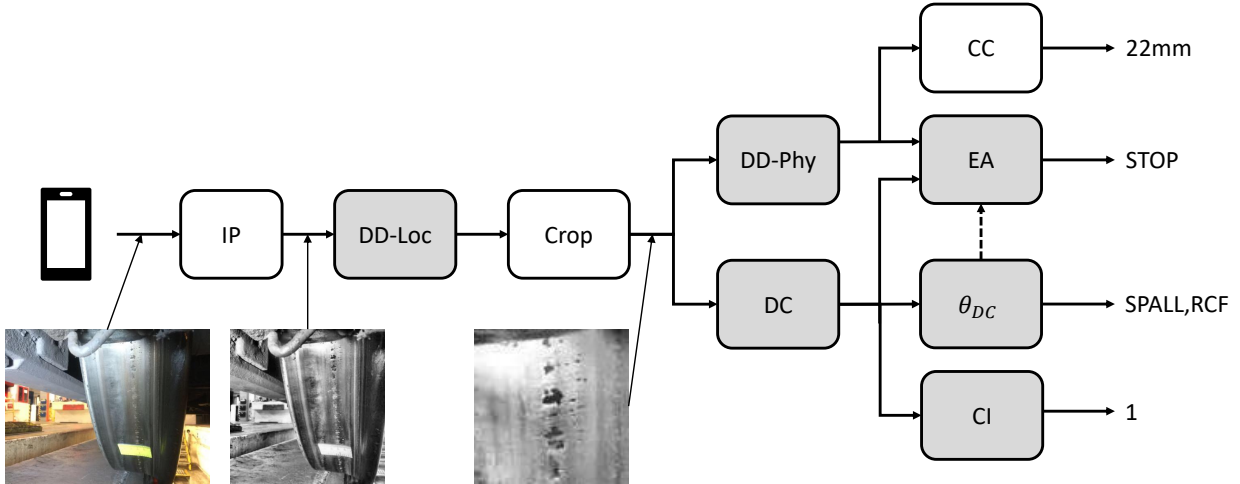


Figure 2. Wheel tread defect diagnosis framework. The main data-driven modules are: Defect Detection (DD-Loc and DD-Phy), Defect Classification (DC and θ_{DC}), Engineering Assessment (EA), and Confidence Index (CI). These are highlighted in shade. The auxiliary modules are: Image Processing (IP), Cropping, and the Circumference Calculation (CC). These are shown in white.

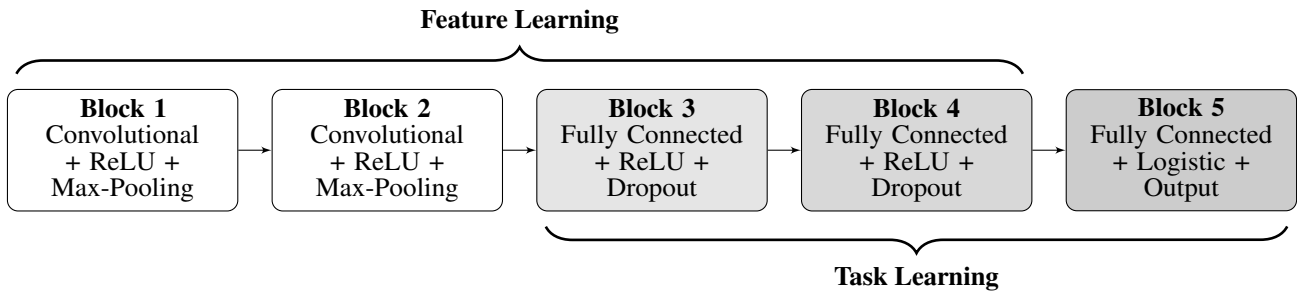


Figure 3. Functional blocks of the proposed versatile unified CNN, each of them containing a layer of learnable weights, an element-wise non-linearity with the ReLU activation function, and a layer of regularization. The Feature Learning blocks are displayed with a white background, whereas the Task Learning blocks have a light shade, showing the transition from the input data to the desired output result.

ear Unit (ReLU) as the non-linear activation function to train faster (Nair, V., and Hinton, G. E., 2010) and avoid the vanishing gradient problem (Glorot, X., Bordes, A., and Bengio, Y., 2011), and Dropout (i.e., random neuron deactivation) to preclude the co-adaptation of the feature detectors (Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R., 2012) and prevent overfitting (Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., 2014). Additionally, a subsampling overlap with a minimum stride of 1 px in a max-pooling step is considered to merge features, increase the robustness to noise, and improve the generalization. In summary, the basic building block of the proposed CNN combines a layer of adjustable weights like the convolutional filters or the fully dense connections, a non-linear rectification transformation (i.e., always positive neuron output values), and a layer of regularization with max-pooling or dropout. The idea of using a block of layers as a structural unit is gaining popularity (Khan, A., Sohail, A.,

Zahoor, U., and Qureshi, A. S., 2020), and therefore this approach is aligned with the latest trends in CNN architecture design. Figure 3 shows this layout, clearly identifying the two learning stages: the features and the task, which are described as follows.

2.4.2. Feature Learning

The Feature Learning stage discovers the degradation-relevant traits in the pictures through a chain of non-linear convolutional and pooling operations, which initially learns simple shapes like curves and straight lines, and then combines these motifs to progressively create more complex and invariant compositions in a higher level of abstraction (Mahendran, A., and Vedaldi, A., 2015), just like many natural signals in visual neuroscience (LeCun, Y. and Bengio, Y., and Hinton, G. E., 2015). It is to note that in the proposed CNN design, no padding is used because there is no useful information in the borders of these images, which always display the defects

in the central region. Once the system has been trained, the adjusted weights of the initial layers (i.e., the image filters) may then be reused throughout the tasks (Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T., 2013; Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S., 2014), which are learned in the following fully-connected layers. The next section delves into the details of this upcoming step.

2.4.3. Task Learning

The Task Learning step is acquired with the remaining fully-connected layer blocks that follow the convolutional blocks, see Figure 3. The non-linear learning capacity is guaranteed with this multilayer structure and the ReLU activation function. It is to note, though, that the last fully-connected block features a logistic sigmoid function, bounding the task dependent variable outputs between 0 and 1.

For the Defect Detection regression tasks (DD-Loc and DD-Phy), a maximum-value normalization step is performed taking into account the picture pixel size for the location task, and a reasonably large physical size for the measurement task. For these targets, a minimum squared error (MSE) training cost is used, which aims to reduce the real-valued prediction residuals.

For the Defect Classification task (DC), a binary class vector is used because the target degradation picture may have many labels (i.e., multiple defects on the same wheel tread). In this case, the cost function of use is the binary cross-entropy, so that each dimension of the output represents the posterior probability of the defect-class membership. This corresponds to the effective deployment of many logistic regressions following the one-vs-all multiclass strategy. Given that the defects are not mutually exclusive, the learning feedback will be shared among the intermediate layers. Finally, a heuristic decision rule based on a threshold is used to discretize the output: a defect class is selected if its predicted probability is over this minimum probability limit.

2.4.4. Feature/Task Embedding

This contribution states that the first two convolutional blocks are mainly meant to deal with the feature learning phase, and the three fully connected blocks that follow mostly learn the task at hand, see Figure 3 for the design diagram that shows the transition between the two stages. This feature/task integration is motivated by the local feature transfer aspect in the convolutional filters (Oquab, M., Bottou, L., Laptev, I., and Sivic, J., 2014), which can detect a particular pattern all over the picture, a characteristic that dense layers do not exhibit due to their rigidity. As it is, the proposed system learns a non-linear but rather shallow set of features, and a deep set of functional task operations. Nonetheless, the boundary between these two objectives in the network is not clear. The same

solution could have been equally described as a profoundly intricate feature learner with four blocks — two convolutional and two fully-connected — and a very shallow linear task learner with only one dense block, which is perhaps the generally adopted CNN functional interpretation. The obtained results would have been the same, especially if the different CNN's are freshly trained or the parameters are reused only for initialization pretraining purposes, but their interpretation would be different.

This work puts forward the contention that the task-specific learned knowledge is effectively embedded in the intermediate fully-connected hidden layers, as their large expressiveness supports this capacity (over 8 million tunable weights for this approach), see Table 2 for a detailed description of the system parameters. Although it has been pointed out that the hidden units may learn similar representations that converge to analogous features across the tasks (Kornblith, S., Norouzi, M., Lee, H., and Hinton, G., 2019), these layers may also experience some optimization difficulties (Yosinski, J., Clune, J., Bengio, Y., and Lipson, H., 2014) (i.e., layers FC3 and FC4). In this last cited reference it is documented that the transferability of features decreases as the distance between the base task and target task increases, thus supporting the rigid task-specific learned knowledge, and limiting the extent of their parameter reuse. This work suggests that only the first two convolutional blocks may be inherited in a different task and all the intermediate dense layers are to be retrained for each different objective.

2.5. Performance Evaluation

Different key performance indicators are used to evaluate the operation of the task-driven CNN approaches. The regression objectives are assessed with the variability of the resulting error distribution for a given confidence interval. This figure is indicative of the amount of epistemic uncertainty. For the classification task, the overall system performance is obtained with the macro-averaged accuracy, precision, and recall metrics (Duda, R. O., Hart, P. E., and Stork, D. G., 2001). These values represent the rate of good classifications, and the penalties that false alarms and missed defects introduce.

In the scenarios where the same dataset is used for learning and evaluation, the performance values are generally estimated with Monte Carlo cross-validation (Dubitzky, W., Granzow, M., and Berrar, D., 2007). Specifically, four rounds of repeated random subsampling are applied on a stratified set of defect types with a train/test split rate of 80/20 (%), which should yields an error sample size over 1k instances that is sufficient to reliably conduct the statistical calculations. In the scenarios where the working dataset is too small for applying this approach, then a leave-one-out cross-validation strategy is pursued. Finally, in the scenarios where both datasets are used, the MRO data is used for training, and the ENG data is

Table 2. CNN parameter chart. The Dropout layers feature a probability of 0.1, and the OR or OC represent the regression or the classification output.

Block	Layer ID	Type	Filter	Stride	Amount	Units	Activation	Parameters
1	C1	Conv2D	(5,5,1)	(1,1)	6	(71,71,6)	ReLU	156
1	P1	Max Pool	(2,2)	(1,1)		(70,70,6)	Linear	0
2	C2	Conv2D	(5,5,6)	(1,1)	16	(66,66,16)	ReLU	2416
2	P2	Max Pool	(2,2)	(1,1)		(65,65,16)	Linear	0
3	FC3	Dense				120	ReLU	8112120
3	D3	Dropout				120	Linear	0
4	FC4	Dense				84	ReLU	10164
4	D4	Dropout				84	Linear	0
5	OR	Dense				2	Logistic	170
5	OC	Dense				6	Logistic	510

held out for testing.

2.6. Development and Industrialization

The Machine Learning research is entirely conducted with the Python3 programming language and its data science ecosystem environment for PHM (Rezaeianjouybari & Shang, 2020), mainly led by NumPy, Scikit-learn and SciPy. For the image processing tasks, OpenCV and scikit-image are also used. Finally, the intensive computations that Deep Learning entails are carried out by TensorFlow2 (Guo, Q., Chen, S., Xie, X., Ma, L., Hu, Q., Liu, H., Liu, Y., Zhao, J., and Li, X., 2019).

The industrialization of the proposed solution for creating a minimum-viable product leverages the latest developments of the open-source big data ecosystem (Cui, Y., Kara, S., and Chan, K. C., 2020). The full architecture stack is running on top of a cluster of machines managed by Kubernetes, a well-proven system to automate, scale and ensure high availability of computer applications. Kubernetes has been increasingly used in the field of machine learning over the past years (Aji, I. P., and Kusuma, G. P., 2020; Wu, C., Haihong, E., and Song, M., 2020). It is divided into four layers: (A) the data layer stores all the data used by the product; (B) the flow layer orchestrates and schedules the “hand-to-hand” transfer of data between the different applications; (C) the application layer centralizes all the “business-value” functions performed by the wheel tread defect diagnosis framework presented in this paper, and (D) the presentation layer contains the user app. The technologies used for each layer, illustrated in Figure 4, are described as follows:

Data Layer PostgreSQL (Juba, S., and Volkov, A., 2019) is used to store the application data such as users, passwords and computation results. It is a well-proven tool with a very powerful query engine. It is used jointly with PostgREST application that creates a REST API on top of PostgreSQL and avoids direct connections which are risky in terms of cybersecurity. MinIO cloud storage (Johnston, C., 2020) is used to upload, store and download the images. It is based on Amazon S3 technology which is

able to handle multiple large binary files downloads and uploads simultaneously without any loss of performance.

Flow Layer Apache NiFi (Chanthakit, S., Keeratiwintakorn, P., and Rattanapoka, C., 2019) is used to orchestrate back and forth the delivery of data between the application and the data layers. It provides a very user-friendly web interface with multiple types of functional blocks (so-called processors) that one can organize and connect together to create more complex flows. One can then follow the traces of the processing path directly in the web interface, which is very practical to monitor the progress. The underlying Kubernetes allows NiFi to run a single flow in a cluster of multiple machines at the same time, thus ensuring the availability of the product.

Application Layer OpenFaaS (Balla, D., Maliosz, M., and Simon, C., 2020) is used to expose the Python scripts for the wheel tread defect diagnosis as a web service executable through a HTTPS request. First the Python scripts and models are encapsulated into a Docker image that is pushed to the OpenFaaS registry. Then OpenFaaS manages the deployment of the Docker image and the routing of requests. OpenFaaS is also increasingly used in the field of Machine Learning (Jang, R-Y., Lee, R., Park, M.-W., and Lee, S.-H., 2020). The underlying Kubernetes allows OpenFaaS to automatically scale up the number of deployed Docker images to smartly adapt the computational power to the actual quantity of requests.

Presentation Layer The Ionic (Yusuf, S., 2016) software development kit is used to develop the user app. The user interface is built as a Progressive Web App (PWA) using the Angular framework jointly with web technologies such as CSS and HTML5. The use of PWA technology allows the mobile app to run both on mobile and web devices (Biørn-Hansen, A., Majchrzak, T. A., and Grønli, T.-M., 2017). The app communicates through classical HTTPS GET and POST requests: with MinIO to post the images and with PostgREST API to get app parameters and computation results.

The main use-case scenario, presented Figure 4, is the follow-

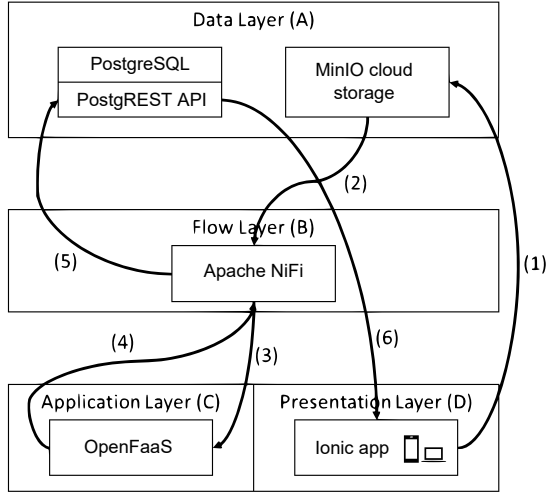


Figure 4. Industrialized architecture of the wheel tread diagnosis framework: four-layer stack with main use-case scenario.

ing: (1) a new image is posted by the maintainer from the user app to MinIO, (2) NiFi takes the image from MinIO, (3) NiFi posts the image to the OpenFaaS gateway to execute the wheel tread defect diagnosis function on its content, (4) OpenFaaS responds to the request with the results of the computation, (5) NiFi inserts the results into the PostgreSQL database through the PostgREST API, and (6) computation results are retrieved by the app and presented to the maintainer (and to the engineer) in the user interface according to the usage profile. The impact of these results on the maintenance business are presented in the following section.

3. RESULTS

This section details the results of the proposed CNN approach to the different specialized tasks to diagnose wheel tread defects and estimates their expected performance.

3.1. Defect Location Performance (DD-Loc)

The defect location module is developed with the MRO dataset. Figure 5 shows the location prediction error distribution scored as the difference between the X and Y coordinates indistinctly. This result shows that the prediction error is centered around the target because there is no bias toward the left/right or up/down. The uncertainty is of 9.5 px, which corresponds to 12.66% of the image size.

3.2. Physical Size Performance (DD-Phy)

The physical size prediction module is also developed with the MRO dataset. Figure 6 shows the error distribution scored as the difference between the width and the height indistinctly. This result shows that the error is sharply centered around the target. The uncertainty of the prediction is of 6.2 mm.

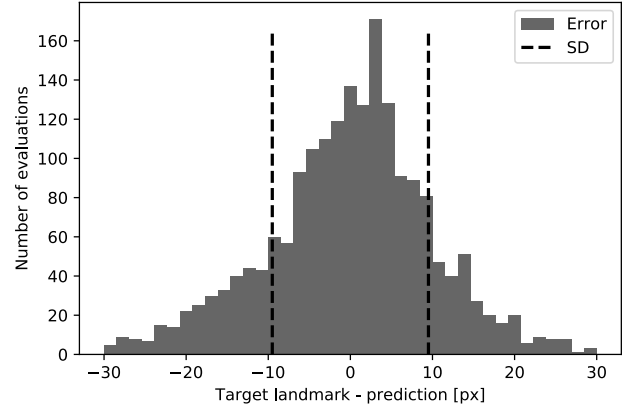


Figure 5. Histogram of the defect location prediction error. The 68% confidence interval SD (i.e., 1 standard deviation under the normality assumption) indicates the uncertainty.

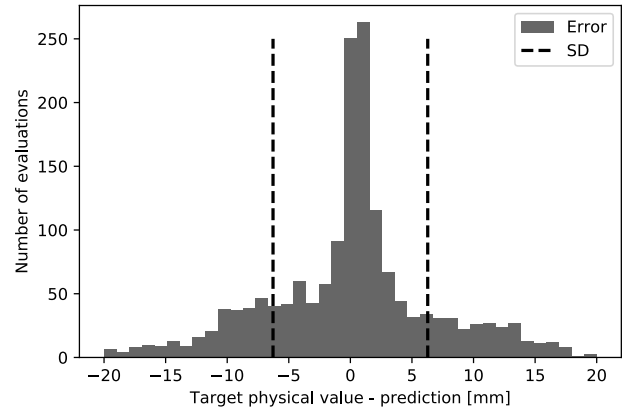


Figure 6. Histogram of the physical size prediction error. The 68% confidence interval SD (i.e., 1 standard deviation under the normality assumption) indicates the uncertainty.

3.3. Defect Classification Performance (DC, θ_{DC})

The classification module that scores the defect type membership probabilities (DC) is trained with the MRO dataset, and the threshold module that discretizes the result (θ_{DC}) is adjusted with the ENG dataset. Figure 7 shows the resulting classification metrics. Note that two potential work points can be identified in the diagram. Their characteristics are described as follows:

- Conservative work point (CWP, $\theta_{DC} = 0.35$): minimize false negatives. With a lower threshold the system yields many potential failure candidates so that the risk of missing a problem is kept low, which is especially important from a safety perspective. The accuracy is higher (0.75) for this configuration.
- Eager work point (EWP, $\theta_{DC} = 0.7$): minimize false alarms. With a higher threshold the system yields fewer potential failure candidates so the system increases its precision (around 0.3). In this configuration, the system is-

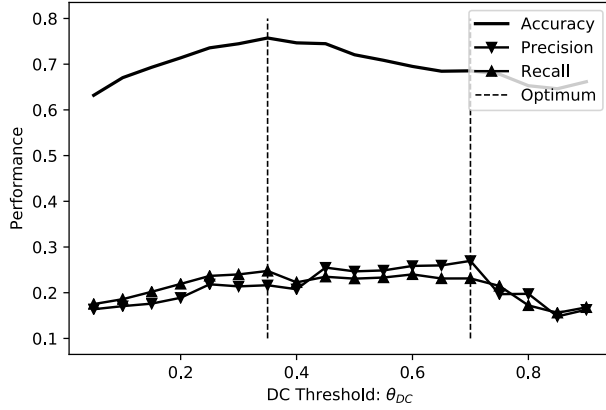


Figure 7. Macro-averaged Defect Classification metrics: accuracy, precision, and recall. Results are shown along with the probability discretization threshold θ_{DC} . Two potential work points are identified.

sues “defect absence” labels whenever all the defect-type probabilities are low, thus enabling it to detect anomalies in compliance with the ISO 13374 international standard (ISO, 2003), i.e., operating as a dichotomic function.

Figure 7 is a kind of Receiver-Operating Characteristic curve, showing more than two key performance indicators. Note that if the threshold θ_{DC} that operates on the vector of defect type probabilities is raised even further (beyond the Eager Work Point), the system is unable to raise any alarm as the required minimum probability values get close to 1.0, and therefore the precision and recall classification metrics drop because they both depend directly on the True Positives of the confusion matrix. Their expected “inverse” behavior is clearly observed at $\theta_{DC} = 0.45$, when the two curves cross. At that point, the system weighs equally the effect of False Positives and False Negatives. In terms of business impact, the priority criteria of the customer ultimately lead the performance tuning process.

Also note that the accuracy performance indicator is not reliable in this imbalanced data scenario, as the system might be biased toward the majority defect type (i.e., RCF), so further operational context is necessary for the evaluation. In the next section, these additional criteria are considered to give a better view of the actual expectations that this proposal provides.

3.4. Engineering Assessment Performance (EA)

This is probably the most decisive module of the system because it provides the actionable feedback in the form of “go - warning - stop” label statements. It is a purely task learning block developed with the ENG dataset. It is built with two of the fully-connected layers of the CNN, yielding a multilayer perceptron architecture. The resulting hidden embedding is arbitrarily set to 10 units (slightly greater than the input dimensionality built with the outputs of the former modules) with Dropout, which will prevent overfitting and ensure that

Table 3. Engineering Assessment performance focused on potential SAF according to different work scenarios: MAC logic rules and conservative/eager work points (CWP/EWP).

Probability	No MAC	MAC + CWP	MAC + EWP
$p(SAF stop)$	0.5	0.34	0.36
$p(SAF warn)$	0.32	0.4	0.32
$p(stop)$	0.08	0.96	0.47
$p(warn)$	0.92	0.04	0.53
$p(SAF)$	0.33	0.34	0.34
$p(SAF; ENG)$	0.37	0.98	0.64

the network automatically finds its optimum expressiveness. In addition, this EA module may eventually apply a series of logical rules known as the minimum acceptance criteria (MAC), which are conservative in nature, to guarantee that certain critical limits are never exceeded.

For the design of this module, its performance in the following three configurations is taken into consideration: no MAC rules, MAC rules with the conservative work point, and MAC rules with the eager work point. Table 3 shows the performance results in probabilistic terms derived from the confusion matrices, and focusing on the potential service-affecting failures (SAF), which are the critical situations identified by the engineering office (i.e., a “stop” label in the ground truth).

This analysis clearly shows the different operating modes: the purely data-driven scenario (i.e., no MAC) is strongly biased toward issuing warning results (just like the majority of the dataset), the conservative scenario is strongly biased toward raising alarms, and the eager scenario is balanced. However, the probability of actually detecting the SAF, which is calculated with the law of total probability, see Eq. (1), is almost the same in all scenarios. Note that the system does not report any “go” result, which may be reasonable because the maintenance staff only take pictures if they suspect the presence of an incipient defect.

$$p(SAF) = \frac{\sum p(SAF|diagnosis) \cdot p(diagnosis)}{p(SAF|stop) \cdot p(stop) + (SAF|warn) \cdot p(warn)} \quad (1)$$

In the light of this inconclusive result where all the approaches yield a probability around 0.34 to detect the potential SAF, the contribution of the engineering team will be determining to break the tie.

3.4.1. Engineering Verification and Validation

Whenever the engineering team checks a picture, it always detects the potential SAF. At present, the engineers manually review the whole stream of images, which takes a lot of person-hour resources and this workload may hinder the completion of other activities. To add value to the overall maintenance pro-

cedure reducing the weight of this engineering validation task, this work proposes that only the pictures automatically rated with the “stop” diagnosis are to be manually checked by the engineering office. In consequence, $p(SAF|stop; ENG) = 1$, and the probability of detecting the SAF increases in different degrees according to the given design strategy. The bottom row of Table 3 shows the impact of this new criterion. These results indicate that with the eager approach (along with the MAC rules) the engineering team can reduce its current workload more than 50%, and retain a SAF detection rate of 64%. This is regarded as the optimum trade-off between the complete manual workload and the complete automated approach, potentially resulting in the best pay off for the adoption of the proposed system.

3.5. Confidence Index Performance (CI)

The Confidence Index informs that the system is self aware of the reliability of its predictions. This indicator is developed with the ENG dataset through a heuristic function that determines the result of this quality test. This function operates on the vector of probabilities of the preceding DC module, and applies an Active Learning approach known as an “acquisition function” that determines the degree of uncertainty in the classification (for all the defect types D) through its entropy (Settles, B., 2010). The resulting value is finally scored against a maximum threshold θ_{CI} to obtain the binary-valued CI, shown in Eq. (2).

$$CI = \left(- \sum_{\forall d \in D} p(d) \cdot \log(p(d)) \right) < \theta_{CI} \quad (2)$$

Figure 8 displays the distribution of the DC entropy for the ENG data, related to their diagnosis labels. Note that for all the instances that display an entropy lower than $\theta_{CI} = 1.2$, the rate of the “stop” diagnostic (i.e., the potential SAF) over the other labels in each bin is considerably greater than the rate over the entropy value of 1.2. Thus, this leads to the conclusion that $\theta_{CI} = 1.2$ is the adequate threshold for the Confidence Index.

The engineering office capitalizes the maintenance expertise, understands the limitations of the proposed CNN system, and the CI indicator can be used to drive their decisions, among other functional criteria. The following section is dedicated to this latter point.

4. DISCUSSION

This section elaborates upon the contextual behavior of the wheel tread defect diagnosis system described in this work.

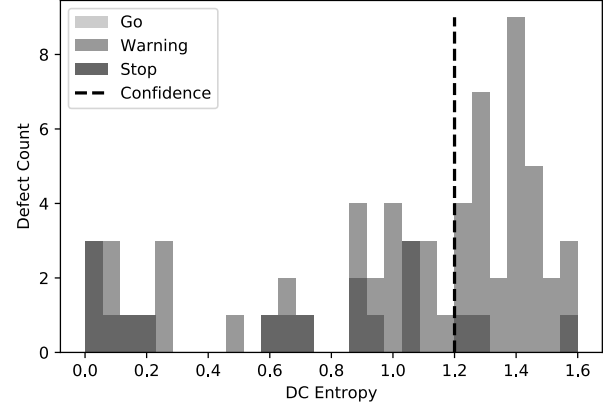


Figure 8. Histogram of the DC entropy for the ENG data with respect to their diagnosis labels.

4.1. Understanding the Learned CNN System

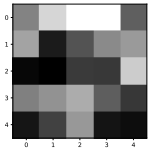
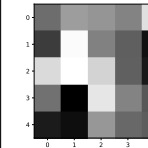
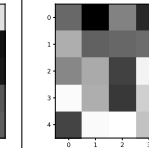
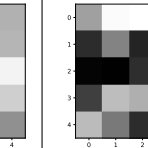
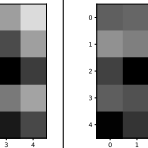
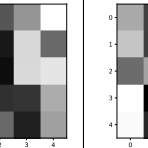
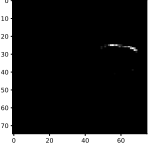
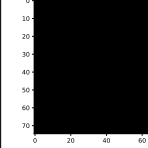
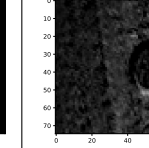
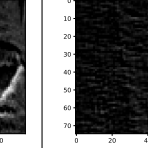
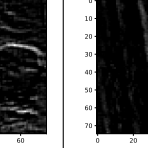
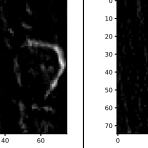
In this multi-label setting where many defects may be represented in the same image, learning the templates for arrangements of objects becomes rapidly intractable because of the combinatorial explosion in the number of features to be stored (Ricci, M., Kim, J., and Serre, T., 2018). One of the main criticisms generally attributed to neural networks, and thus to CNN’s in particular, is their lack of interpretability or explainability, what is also known as a “black box” interface behavior (Fong, R., and Vedaldi, A., 2017). The following two sections delve into the internal working details of the learned CNN system in order to shed some light into their cumbersome operations, revealing the desirable properties of compositionality and class discrimination that CNN’s are expected to exhibit (Zeiler, M. D., and Fergus, R., 2014).

4.1.1. Image Filters

A CNN is fundamentally characterized by the adapted design of its filters, which get convoluted with the input image in order to highlight interesting patterns, just like the human visual system (Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B., 2017). In a sense, these filters are like templates that match specific motifs in the pictures, especially the ones found in the first layer of a vision system (Erhan, D., Bengio, Y., Courville, A., and Vincent, P., 2009), where the receptive field, i.e., the size of the region in the input that produces the feature, is minimum and corresponds to the size of the filter (Le, H., and Borji, A., 2017). It is widely accepted that these first functions learn edge-detecting Gabor filters, i.e., linear functions used for texture analysis that highlight a specific frequency content in a specific selective direction. Therefore, analyzing them at the pixel level reveals relevant information about the captured knowledge (Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W., 2015).

The outputs of the filters correspond to specific locations of

Table 4. First layer of learned image filters (C1) and their impact on a sample image showing two defects (SPALL and RCF).

Filter						
Output						
Function	Up curve	(None)	Down curve	Up/down curve	Vertical line, right curve	Vertical line, right/left curve
Defect	SPALL	(None)	SPALL	SPALL	SPALL, RCF	SPALL, RCF

interest whenever their activation is high, thus creating a spatial feature detector (Zeiler, M. D., and Fergus, R., 2014). And given that these patterns can be observed in any place around the picture, their dependence on individual units is reduced, thereby improving the network generalization performance (Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., and Botvinick, M., 2018). Table 4 shows the first filters that the system has learned (i.e., layer C1) and the impact of their design on a sample image that contains two tread defects (SPALL and RCF). As it can be seen, each of the six input filters learns a particular detail of the degradation: some filters learn curves, others learn straight lines, and even two of them learn both features, illustrating the multifaceted character of the related neurons (Nguyen, A., Yosinski, J., and Clune, J., 2016). In most cases, their output can then be directly related to a specific type of defect, which gives them a kind of latent representation aligned with human-interpretable semantic concepts (Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A., 2017). However, it is the entire space of activations, rather than the individual units, that contains the bulk of the semantic information (Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R., 2013). Finally, all these features get blended into the layers that follow to accomplish some task-driven goal.

4.1.2. Defect Manifold

This section evaluates the separability of the spatial distribution of the defects in the latent space (Chen, Z., Bei, Y., and Rudin, C., 2020). To see how the CNN architecture internally discriminates the data and manages the inter-defect knowledge (Mahendran, A., and Vedaldi, A., 2015; Simonyan, K., Vedaldi, A., and Zisserman, A., 2013; Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., 2015) as well as the intra-defect knowledge through a hierarchical and compositional pipeline (Wei, D., Zhou, B., Torralba, A., and Freeman, W. T., 2015), Figure 9 shows the scattering of the instances on

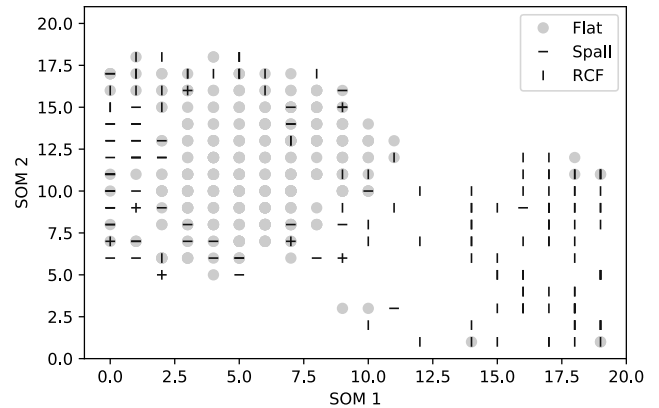


Figure 9. Self Organizing Map of the FC4 layer embedding for the defect classification task showing the main defect types: flat, spall, and RCF.

the penultimate adjustable layer FC4 for the defect classification task using a Self Organizing Map (SOM) (Kohonen, T., 1990). The SOM is an unsupervised non-linear transformation technique based on competitive learning that produces a discretized representation of the data preserving its topological properties, i.e., its similarity clusters. In PHM it has been used for anomaly detection and fault location purposes (Tian, J., Azarian, M. H., and Pecht, M., 2014; Zhao, W., Siegel, D., Lee, J., and Su, L., 2013), also in railway systems (Alessi, A., La-Cascia, P., Lamoureux, B., Pugnali, M., and Dersin, P., 2016).

In the scenario presented in this work, the SOM shows how the CNN learns to separate the three major defect prototypes: RCF, flat, and spall. In particular, it can be observed that the system learns to differentiate straight-line patterns (e.g., RCF), which are clustered to the right, from rounded patterns (i.e., spall and flat), which are clustered to the left. In this latter categorization, the overlap illustrates that the curvy-type

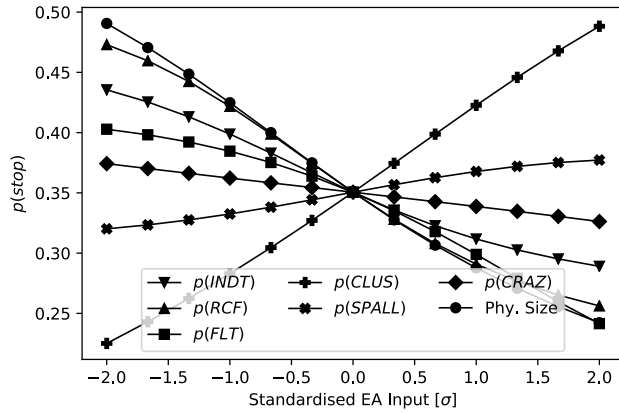


Figure 10. Engineering Assessment sensitivity analysis through the profile method for the “stop” diagnosis statement. The inputs have been standardized.

patterns seem to follow a hierarchical structure over the defects (Alsallakh, B., Jourabloo, A., Ye, M., Liu, X., and Ren, L., 2018). Note that only the last layer of the CNN deals with this manifold representation, and the eventual classification thus needs to be attained with linear discriminators, which seems to be adequate based on the lower-dimensional representation provided by the SOM. However, a proper expressiveness analysis with an additional hidden layer, thus creating a multi-layer perceptron, could be a more general solution (Simard, P. Y., Steinkraus, D., and Platt, J. C., 2003), following the universal approximation theorem for neural networks (Cybenko, G., 1989; Pinkus, A., 1999).

4.2. Engineering Assessment Sensitivity

The Engineering Assessment module is arguably the most critical point in the system because it provides the actionable feedback to the maintainer. To understand its inner working mechanism through the impact of the input variables (i.e., the defect type probabilities and the physical size of the defect) on the output diagnosis, Figure 10 displays the result of a sensitivity analysis based on the profile method (Shojaeefard, M. H., Akbari, M. Tahani, M., and Farhani, F., 2013) for the critical “stop” diagnostic.

Assuming that the importance of a variable is driven by the dynamic range of the output, it is shown that the physical size, the RCF, and the CLUS probabilities lead this ranking. In addition, the physical size and the RCF probability variables are strongly negatively correlated with the “stop” probability diagnosis. The RCF is a type of defect that by itself does not directly halt the railway service, so this negative relationship makes sense. The physical defect, however, does not reasonably follow this criterion, but its impact is highly correlated with the RCF and this is what the EA module has ultimately learned. Finally, the CLUS probability is strongly positively correlated with the diagnosis, which makes perfect sense be-

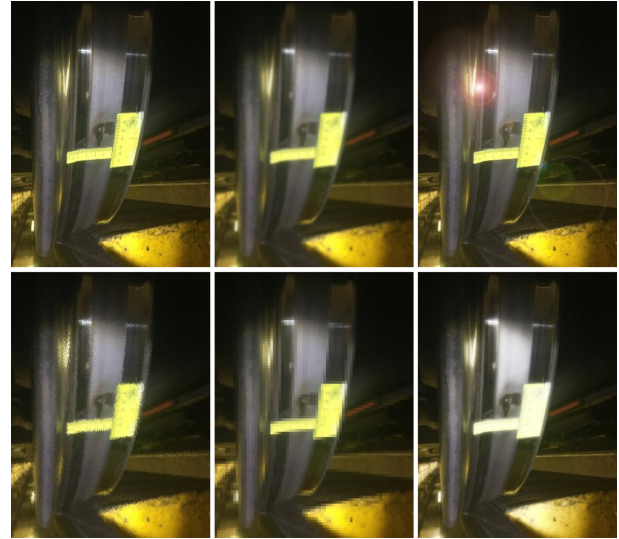


Figure 11. Examples of image modifications: normal picture, blur, glare, noise, pixelation, and shine.

cause this is a critical defect type.

4.3. Robustness to Feature Corruption

Machine learning solutions, including neural networks and Deep Learning, may exhibit unexpected instability on simple perturbations. Therefore, they are at risk of being tricked by adversarial instances, which are intentionally corrupted data that lead the system to output incorrect results with high confidence (Goodfellow, I. J., Shlens, J., and Szegedy, C., 2015). Moreover, image processing applications are especially targeted by these attacks because some of these small perturbations are difficult to detect as they exploit edge cases. Methods such as histogram equalization, see Section 2.2.1, are helpful to prevent them (Hendrycks, D., and Dietterich, T., 2019), but careful attention is needed because cybersecurity in railways is an area that has attracted a lot of interest recently due to an increasing number of denial-of-service attacks (Masson, É., and Gransart, C., 2017).

A useful approach to build a defense against these adversarial attacks is to construct a predictor that is robust to the deletion of features at test time (Globerson & Roweis, 2009). In this sense, the Engineering Assessment module already features a Dropout layer after the embedding, see Section 3.4. In the defect diagnosis scenario based on smartphone pictures presented in this work, the proposed system should be robust to artificial image modifications that could be used in an adversarial attack, including effects like blurring, flash glare, etc. Figure 11 shows some typical examples of these kind of tweaks, and Table 5 evaluates their impact on the final diagnosis.

This analysis of feature perturbations shows that the proposed system exhibits a fairly good overall robustness to potential im-

Table 5. Image modifications and their impact to the proposed defect diagnosis system.

Filter/Effect	Size (mm)	Defects	Diagnosis	CI
Normal	55	FLT	Stop	1
Blur	54	None	Warning	0
Glare	61	FLT	Stop	1
Noise	53	None	Warning	1
Pixelation	68	FLT	Stop	0
Shine	53	FLT	Stop	1

age corruptions, including the common shine and glare effects produced by the flash. Robustness to pixelation also indicates that the resolution of the smartphone camera is sufficient. Nevertheless, the blur and noise perturbations cause the system to fail, as a warning signal is issued instead of the expected “stop” statement. These situations thus need to be avoided through the recommendation of taking still photographs in a dust-free environment.

4.4. Pragmatic Project Management

The development of an industrial Deep Learning solution entails having to deal with many different components, and this leaves the door open to many different potential approaches. On the data acquisition stage, a Computer Vision engineer will probably argue that the system improvement lies on the quality of the pictures, and these smartphone images do have focus issues, uneven lighting conditions, different distances to the wheelset defect of interest, etc. However, when the input pictures are taken at different scales, the CNN will extract features at different scales (He, K., Zhang, X., Ren, S., and Sun, J., 2015), so these variations should not be the primary point of concern. Moreover, it has been shown that the resolution of the camera (leading to a pixelation effect) does not critically impact the diagnosis.

What has been observed is the tight dependence on labeled data to develop such a system. The annotation process is tedious, and fatigue builds up after some time. In this work, a standalone computer application has been developed to iterate the dataset and record the expertise, which has been provided by one single expert per instance. A minimum inter-annotator agreement rate is not strictly necessary for a feasible tagging of maintenance data (Hastings, E. M., Sexton, T., Brundage, M. P., and Hodkiewicz, M., 2019). However, further progress in this line should be provided by unsupervised or semi-supervised approaches, which reduce the amount of repetitive human effort (Bengio, Y., 2009), like the Meta Pseudo Labels approach (Pham, H., Dai, Z., Xie, Q., Luong, M.-T., and Le, Q. V., 2020), where a teacher network is trained to generate pseudo labels on unlabeled data to train a student network, and adapts with the performance of the student network on the labeled dataset. Additionally, a strategy to reduce the bias in the data (Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J., 2019) and the noise in the labels (Lee, K.-H., He, X.,

Zhang, L., and Yang, L., 2018) should also be explored.

On the learning stage, the proposed CNN design displays zero bias error throughout the different modules, and any tweak beyond this neural design has led to the appearance of some average loss (keeping the same uncertainty). Therefore, as it is, the described approach shows an optimum complexity for this defect diagnosis problem, despite the obtained results are far from perfect. However, it is not clear how a different architecture might be of help in this scenario. There are some approaches that suggest using smaller convolutional filters (3x3) along with a network depth of 16 to 19 layers (Simonyan, K., and Zisserman, A., 2015), keeping a constant computational budget for the industrialization (Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., 2015), using kernels within the convolution function (Ammann, O., Michau, G., and Fink, O., 2020), or dropping the pooling layers due to their seldom attributed destructive role (Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M., 2015). The Deep Learning field is in full blossom at present, and potentially many different solution approaches to the problem will be developed, so further research is required to get an optimal solution and to settle into the plateau of general productivity. Ultimately, the obtained solution as it is could be used to train a new generation of networks in a self-distillation manner and push the test performance a bit further (Zhang, L., Song, J., Gao, A., Chen, J., Bao, J., and Ma, K., 2019).

Alternatively, the current technology may also be used with a different perspective: instead of the proposed modular approach, a truly multitask environment could also be explored, because a single network can manage to do classification and regression tasks concurrently (Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., 2015; Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y., 2013). What is more, the application of a CNN at multiple locations in a sliding window fashion (instead of the full image input) has also been reported to be successful (Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y., 2013; Oquab, M., Bottou, L., Laptev, I., and Sivic, J., 2014). In addition, the domain transfer between a rich image environment like ImageNet and the defect problem at hand may also be of help to learn better feature representations and improve the system generalization (Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A., 2020; Kornblith, S., Shlens, J., and Le, Q. V., 2019). Furthermore, the consideration of synthetic data including adversarial images (Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A., 2019), which is a popular approach to train Deep Learning models for Computer Vision (Nikolenko, S. I., 2019), is a useful resource to enhance the robustness of the system. And in the line of continuous improvement, if the user feedback is included with respect to the presented diagnosis results, the system can also exhibit some sort of enhanced evolution as

new data is processed.

Finally, regarding the industrialization, the main limitation of the proposed architecture is that the Python models used for the whole machine learning process are included in a Docker container image that is almost immutable, hence dynamic updates of the models are cumbersome. An upgrade to the proposed solution could include a kind of model registry that is periodically called to download new versions of the models, which would be developed by the data scientist in the loop following a continuous improvement procedure driven by the Return on Experience of the product, including new features, bug fixes, patches motivated by incorrect predictions, etc.

5. CONCLUSION

The detection of railway wheel tread defects on raster picture data is a daunting task that involves many different levels of analysis. This paper presents an integrated solution based on many Convolutional Neural Networks that locate the damaged areas in the images, estimate the physical size of the shown defects, and assess their type and severity. This proposal describes a task-division approach that helps understand the caveats and pitfalls of the predictive value chain. The results indicate that almost half the current engineering effort dedicated to manually checking the potential issues can now be automated, thus reducing the lead time to take a timely maintenance action, and ultimately optimizing the activities of the workforce.

The future work that is currently envisaged may further deal with the following topics:

- The explicit consideration of a “good” condition class to better understand the whole image degradation spectrum of the wheel tread defects. Although in the current scenario this is not strictly necessary because the maintenance staff already applies their criteria to take a picture, if this additional assessment was managed as a separate anomaly detection step prior to the described analysis, the whole pipeline would introduce a kind of double-check procedure.
- The collection of actual feedback from the field and the evaluation of the value added by the diagnosis. The Appendix shows some additional examples obtained with the minimum-viable product that is derived from the industrialization of the proposed solution. System interface feedback is also included in the continuous improvement of this online tool.
- The utility expansion to other types of wheels. Despite the proposed solution is tailored to steel railway wheelsets, the same technology can be applied to other types of wheels because CNNs ultimately tend to focus on their texture (Hermann, K. L., Chen, T., and Kornblith, S., 2020). For example, rubber-based tires would display patterns of deflation, punctures, tears or bulges on the

sidewalls, etc.

- In terms of safety, the proposed modular system is advantageous because the EN 50126-1 international railway standard specifies that such systemic hierarchy enables the assessment of subsystem interactions (CENELEC, 2017), and this is a prerequisite to understanding its overall limitations.
- In terms of security in a Deep Learning environment for Computer Vision, further robustness to adversarial images should also be studied, in addition to other cybersecurity considerations. In this sense, technologies like the Digital Twin enables virtual representations of components and systems (Moyné, J., Balta, E. C., Kovalenko, I., Faris, J., Barton, K., and Tilbury, D. M., 2020), which can help detect the presence of anomalous behaviors driven by attacks.

ACKNOWLEDGMENT

We would like to show our gratitude to Fahd Janjua for his help with the data and engineering expertise, Joaquim Serra for his support with the low-level image processing tools, Verónica Fernández for her effort with the Self Organizing Map and the sensitivity analysis, Sergi Bermejo, Guillermo Sospedra and Vicente Fuerte for their management advice, and Alstom’s Innovation Board for funding this project. The contribution of Alexandre Trilla to this research was partially supported by the Government of Catalonia (Generalitat de Catalunya) Grant No. 2020 DI 54.

APPENDIX

Additional examples of actual wheel tread defects along with their diagnostics are shown in Figure 12 and Figure 13.



Figure 12. Picture of a mild spall defect.

REFERENCES

- Aji, I. P., and Kusuma, G. P. (2020). Landmark Classification Service Using Convolutional Neural Network and Ku-

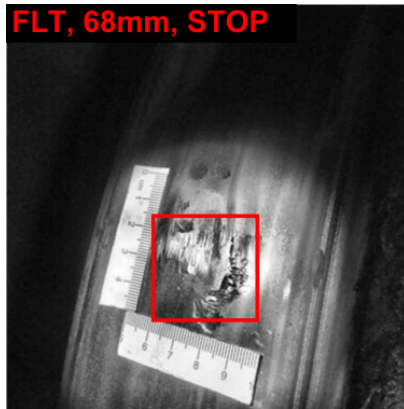


Figure 13. Picture of a critical flat defect.

bernetes. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(3), 2817–2823.

- Alessi, A., La-Cascia, P., Lamoureux, B., Pugnaroni, M., and Dersin, P. (2016). Health Assessment of Railway Turnouts: A Case Study. *Proc. of the Third European Conference of the Prognostics and Health Management Society*, 1–8.
- Alsallakh, B., Jourabloo, A., Ye, M., Liu, X., and Ren, L. (2018). Do Convolutional Neural Networks Learn Class Hierarchy? *IEEE Transactions on Visualization and Computer Graphics*, 4, 1–11.
- Ammann, O., Michau, G., and Fink, O. (2020). Anomaly Detection And Classification In Time Series With Kernel Convolutional Neural Networks. *Proc. of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*, 1–8.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, 10(7), 1–46.
- Balla, D., Maliosz, M., and Simon, C. (2020). Open Source FaaS Performance Aspects. *Proc. of the 43rd International Conference on Telecommunications and Signal Processing*, 358–364.
- Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network Dissection: Quantifying Interpretability of Deep Visual Representations. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 6541–6549.
- Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2(1), 1–71.
- Biørn-Hansen, A., Majchrzak, T. A., and Grønli, T.-M. (2017). Progressive web apps: The possible web-native unifier for mobile development. *Proc. of the International Conference on Web Information Systems and Technologies*, 344–351.
- CENELEC. (2017). *Railway Applications - The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS) - Part 1: Generic RAMS Process* (Tech. Rep. No. 50126-1:2017). European Committee for Electrotechnical Standardization.
- Chanthakit, S., Keeratiwintakorn, P., and Rattanapoka, C. (2019). An IoT System Design with Real-Time Stream Processing and Data Flow Integration. *Research, Invention, and Innovation Congress*, 1–5.
- Chen, Z., Bei, Y., and Rudin, C. (2020). Concept Whitening for Interpretable Image Recognition. *Nature Machine Intelligence*, 2, 772–782.
- Cui, Y., Kara, S., and Chan, K. C. (2020). Manufacturing big data ecosystem: A systematic literature review. *Robotics and computer-integrated Manufacturing*, 62(101861).
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2, 303–314.
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2013). DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *arXiv:1310.1531 [cs.CV]*, 1–10.
- Dubitzky, W., Granzow, M., and Berrar, D. (2007). Fundamentals of data mining in genomics and proteomics. *Springer Science & Business Media*, 178.
- Duda, R. O., Hart, P. E., and Stork, D. G. (Ed.). (2001). *Pattern Classification*. Wiley-Interscience.
- Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152, 184–194.
- Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. (2019). Exploring the Landscape of Spatial Robustness. *Proc. of the 36th International Conference on Machine Learning*, 1–23.
- Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing Higher-Layer Features of a Deep Network. *University of Montreal - Technical Report 1341*, 1–14.
- 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.
- Fong, R., and Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 3429–3437.
- Globerson, T. C. H. S. A., A., & Roweis, S. (2009). An adversarial view of covariate shift and a minimax approach. In S. M. S. A. Quiñero-Candela J. & N. D. Lawrence (Eds.), *Dataset shift in machine learning* (pp. 179–197). Cambridge, Massachusetts: The MIT Press.
- Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proc. of the 14th International*

- Conference on Artificial Intelligence and Statistics*, 315–323.
- Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *Proc. of the International Conference on Learning Representations*, 1–11.
- Guo, G., Peng, J., Yang, K., Xie, L., and Song, W. (2017). Wheel Tread Defects Inspection Based on SVM. *Far East NDT New Technology Application Forum*, 251–253.
- Guo, Q., Chen, S., Xie, X., Ma, L., Hu, Q., Liu, H., Liu, Y., Zhao, J., and Li, X. (2019). An Empirical Study towards Characterizing Deep Learning Development and Deployment across Different Frameworks and Platforms. *Proc. of the IEEE/ACM International Conference on Automated Software Engineering*, 810–822.
- Han, K., Sun, M., Zhou, X., Zhang, G., Dang, H., and Liu, Z. (2017). A new method in wheel hub surface defect detection: Object detection algorithm based on deep learning. *Proc. of the International Conference on Advanced Mechatronic Systems*, 335–338.
- Hastings, E. M., Sexton, T., Brundage, M. P., and Hodkiewicz, M. (2019). Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 1–7.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 1904–1916.
- Hendrycks, D., and Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proc. of the International Conference on Learning Representations*, 1–16.
- Hermann, K. L., Chen, T., and Kornblith, S. (2020). The Origins and Prevalence of Texture Bias in Convolutional Neural Networks. *Proc. of the 34th Conference on Neural Information Processing Systems*, 1–25.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580 [cs.NE]*, 1–18.
- Hyde, P., Ulianov, C., and Defosse, F. (2016). Development and testing of an automatic remote condition monitoring system for train wheels. *IET Intelligent Transport Systems*, 10(1), 32–40.
- Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial Examples Are Not Bugs, They Are Features. *Proc. of the 33rd Conference on Neural Information Processing Systems*, 1–12.
- ISO. (2003). *Condition monitoring and diagnostics of machine systems: Data processing, communication and presentation* (Tech. Rep. No. 13374-1:2003). International Organization for Standardization.
- Jang, R.-Y., Lee, R., Park, M.-W., and Lee, S.-H. (2020). Development of an AI Analysis Service System based on OpenFaaS. *The Journal of the Korea Contents Association*, 20(7), 97–106.
- Jo, J., and Bengio, Y. (2017). Measuring the tendency of CNNs to Learn Surface Statistical Regularities. *arXiv:1711.11561 [cs.LG]*, 1–13.
- Johnston, C. (2020). *Data Lakes*. Berkeley, CA, USA: Apress.
- Juba, S., and Volkov, A. (2019). *Learning PostgreSQL 11: a beginner's guide to building high-performance PostgreSQL database solutions*. Packt Publishing Ltd.
- Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 1–70.
- Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. (2019). Learning Not to Learn: Training Deep Neural Networks with Biased Data. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 9012–9020.
- Kohonen, T. (1990). The Self-Organizing Map. *Proc. of the IEEE*, 78(9), 1464–1480.
- Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of Neural Network Representations Revisited. *Proc. of the International Conference on Machine Learning*, 1–20.
- Kornblith, S., Shlens, J., and Le, Q. V. (2019). Do Better ImageNet Models Transfer Better? *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2661–2671.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
- Krummenacher, G., Ong, C. S., Koller, S., Kobayashi, S., and Buhmann, M. (2018). Wheel Defect Detection With Machine Learning. *IEEE Transactions on Intelligent Transportation Systems*, 19(4), 1176–1187.
- Le, H., and Borji, A. (2017). What are the Receptive, Effective Receptive, and Projective Fields of Neurons in Convolutional Neural Networks? *arXiv:1705.07049 [cs.CV]*, 1–7.
- LeCun, Y. and Bengio, Y., and Hinton, G. E. (2015). Deep Learning. *Nature*, 521, 436–444.
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- Lee, K.-H., He, X., Zhang, L., and Yang, L. (2018). CleanNet: Transfer Learning for Scalable Image Classifier Training With Label Noise. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 5447–5456.
- Ma, K., Vicente, T. F. Y., Samaras, D., Petrucci, M., and Magnus, D. L. (2016). Texture classification for rail surface condition evaluation. *Proc. of the IEEE Winter Conference on Applications of Computer Vision*, 1–9.

- Magel, E., and Kalousek, J. (1996). Identifying and Interpreting Railway Wheel Defects. *Proc. of the International Heavy Haul Association Conference on Freight Car Trucks/Bogies*, 7–20.
- Mahendran, A., and Vedaldi, A. (2015). Understanding Deep Image Representations by Inverting Them. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 5188–5196.
- Masson, É., and Gransart, C. (2017). Cyber Security for Railways - A Huge Challenge - Shift2Rail Perspective. *Proc. of the International Workshop on Communication Technologies for Vehicles. Lecture Notes in Computer Science*, 10222, 97–104.
- Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., and Botvinick, M. (2018). On the importance of single directions for generalization. *Proc. of the International Conference on Learning Representations*, 1–15.
- Moyne, J., Balta, E. C., Kovalenko, I., Faris, J., Barton, K., and Tilbury, D. M. (2020). A Requirements Driven Digital Twin Framework: Specification and Opportunities. *IEEE Access*, 8, 107781–107801.
- Nair, V., and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proc. of the 27th International Conference on Machine Learning*, 1–8.
- Nguyen, A., Yosinski, J., and Clune, J. (2016). Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned by Each Neuron in Deep Neural Networks. *arXiv:1602.03616 [cs.NE]*, 1–23.
- Nikolenko, S. I. (2019). Synthetic Data for Deep Learning. *arXiv:1909.11512 [cs.LG]*, 1–156.
- Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 1717–1724.
- Pham, H., Dai, Z., Xie, Q., Luong, M.-T., and Le, Q. V. (2020). Meta Pseudo Labels. *arXiv:2003.10580 [cs.LG]*, 1–22.
- Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8, 143–195.
- Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B. t. H., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive Histogram Equalization and Its Variations. *Computer Vision, Graphics, and Image Processing*, 39, 355–368.
- Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 512–519.
- Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y. L., Tan, J., Le, Q. V., and Kurakin, A. (2017). Large-Scale Evolution of Image Classifiers. *Proc. of the 34th International Conference on Machine Learning*, 70, 2902–2911.
- Rezaeianjouybari, B., & Shang, Y. (2020). Deep learning for prognostics and health management: State of the art, challenges, and opportunities. *Measurement*, 163(107929).
- Ricci, M., Kim, J., and Serre, T. (2018). Same-different problems strain convolutional neural networks. *arXiv:1802.03390 [cs.CV]*, 1–6.
- Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. (2020). Do Adversarially Robust ImageNet Models Transfer Better? *Proc. of the 34th Conference on Neural Information Processing Systems*, 1–31.
- Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv:1312.6229 [cs.CV]*, 1–16.
- Settles, B. (2010). Active Learning Literature Survey. *University of Wisconsin–Madison - Computer Sciences Technical Report 1648*, 1–67.
- Shang, L., Yang, Q., Wang, J., Li, S., and Lei, W. (2018). Detection of rail surface defects based on CNN image recognition and classification. *Proc. of the International Conference on Advanced Communication Technology*, 45–51.
- Shojaefard, M. H., Akbari, M. Tahani, M., and Farhani, F. (2013). Sensitivity Analysis of the Artificial Neural Network Outputs in Friction Stir Lap Joining of Aluminum to Brass. *Advances in Materials Science and Engineering*, 1–7.
- 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*, 1–6.
- Simonyan, K., and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proc. of the International Conference on Learning Representations*, 1–14.
- Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cv.CV]*, 1–8.
- Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for Simplicity: The All Convolutional Net. *Proc. of the International Conference on Learning Representations*, 1–14.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, 1929–1958.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna,

- Z. (2016). Rethinking the Inception Architecture for Computer Vision. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv:1312.6199 [cs.CV]*, 1–10.
- Tan, M., and Quoc, V. L. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proc. of the International Conference on Machine Learning*, 1–10.
- Tian, J., Azarian, M. H., and Pecht, M. (2014). Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm. *Proc. of the European Conference of the Prognostics and Health Management Society*, 1–9.
- Vickerstaff, A., Bevan, A., and Boyacioglu, P. (2020). Predictive Wheel-rail Management in London Underground: Validation and Verification. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 234(4), 393–404.
- Wei, D., Zhou, B., Torralba, A., and Freeman, W. T. (2015). Understanding Intra-Class Knowledge Inside CNN. *arXiv:1507.02379 [cs.CV]*, 1–7.
- Wu, C., Haihong, E., and Song, M. (2020). An Automatic Artificial Intelligence Training Platform Based on Kubernetes. *Proc. of the 2nd International Conference on Big Data Engineering and Technology*, 58–62.
- Yang, Y., and Xu, Z. (2020). Rethinking the Value of Labels for Improving Class-Imbalanced Learning. *Proc. of the 34th Conference on Neural Information Processing Systems*, 1–21.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How Transferable Are Features in Deep Neural Networks? *Proc. of the 27th International Conference on Neural Information Processing Systems*, 2, 3320–3328.
- Yusuf, S. (2016). *Ionic Framework By Example*. Packt Publishing Ltd.
- Zeiler, M. D., and Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *Proc. of the European Conference on Computer Vision. Lecture Notes in Computer Science*, 8689, 818–833.
- Zhang, J., Guo, Z., Jiao, T., and Wang, M. (2018). Defect Detection of Aluminum Alloy Wheels in Radiography Images Using Adaptive Threshold and Morphological Reconstruction. *Applied Sciences*, 8(2365), 1–12.
- Zhang, L., Song, J., Gao, A., Chen, J., Bao, J., and Ma, K. (2019). Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. *arXiv:1905.08094 [cs.LG]*, 1–10.
- Zhang, W., Zhang, Y., Li, J., Gao, X., and Wang, L. (2014). The defects recognition of wheel tread based on linear CCD. *IEEE Far East Forum on Nondestructive Evaluation/Testing*, 302–307.
- Zhang, Y., Cui, X., Liu, Y., and Yu, B. (2018). Tire Defects Classification Using Convolution Architecture for Fast Feature Embedding. *International Journal of Computational Intelligence Systems*, 11, 1056–1066.
- Zhao, W., Siegel, D., Lee, J., and Su, L. (2013). An Integrated Framework of Drivetrain Degradation Assessment and Fault Localization for Offshore Wind Turbines. *International Journal of Prognostics and Health Management*, 1–13.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Object detectors emerge in Deep Scene CNNs. *Proc. of the International Conference on Learning Representations*, 1–12.

BIOGRAPHIES

Alexandre Trilla graduated from La Salle University of Barcelona with a M.Sc. in Electronics and Telecommunications 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 (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 with artificial neural networks.

John Bob-Manuel is a Mechanical Systems Engineer at Alstom, where he is responsible for the performance of the mechanical systems of London Underground's Northern Line fleet in order to maintain and improve the reliability, availability, and safety of the trains. John obtained his M.Sc. in Advanced Mechanical Engineering from Imperial College London, and received his Bachelor's degree in Aerospace Engineering from Queen Mary University of London. Before joining Alstom, he worked in the oil & gas and aerospace industries, where he led multi-disciplinary engineering teams to design and analyze systems ranging from remote, deep offshore structures to aircraft interior systems. He has a keen interest in using data science to help inform business decisions in specific areas of engineering operations.

Benjamin Lamoureux is a Data Scientist and Predictive Maintenance engineer at Alstom. Dr. Lamoureux has received both a Master's Degree in Mechatronics from the University Pierre & Marie Curie and a Master's Degree in Engineering from Arts & Métiers ParisTech in Paris in 2010. From 2011 to 2014, he performed an industrial Ph.D. work in collaboration between Arts & Métiers ParisTech and SAFRAN Aircraft Engines Villaroche (formerly SAFRAN Snecma). He received his Ph.D. in June 2014. In October 2014, he joined the PHM department of Alstom Saint-Ouen to extend his Ph.D. research

in the railway domain, and he still occupies the same position today. His current research interests are machine learning, statistics, data science and computer science applied to PHM.

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.

5.4 Journal Article 2 (2022)

Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting (2022)

This contribution develops a causal language embedding strategy using a neural encoder, which models the textual entailment of Return On Experience data for bogie diagnosis in a root cause analysis troubleshooting scenario.

This article was first published on July 2022 in the International Journal of Prognostics and Health Management (Trilla, A., Mijatovic, N., and Vilasis-Cardona, X., 2022), and then presented on September 2023 at the 2nd Causal AI Conference in New York City.

Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting

Alexandre Trilla^{1,3}, Nenad Mijatovic², and Xavier Vilasis-Cardona³

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 explores how the causality inference paradigm may be applied to troubleshoot the root causes of failures through language processing and Deep Learning. To do so, the causality hierarchy has been taken for reference: associative, interventional, and retrospective levels of causality have thus been researched within textual data in the form of a failure analysis ontology and a set of written records on Return On Experience. A novel approach to extracting linguistic knowledge has been devised through the joint embedding of two contextualized Bag-Of-Words models, which defines both a probabilistic framework and a distributed representation of the underlying causal semantics. This method has been applied to the maintenance of rolling stock bogies, and the results indicate that the inference of causality has been partially attained with the currently available technical documentation (consensus over 70%). However, there is still some disagreement between root causes and problems that leads to confusion and uncertainty. In consequence, the proposed approach may be used as a strategy to detect lexical imprecision, make writing recommendations in the form of standard reporting guidelines, and ultimately help produce clearer diagnosis materials to increase the safety of the railway service.

1. INTRODUCTION

Natural Language Processing (NLP) provides an effective approach for improving the collection and analysis of text-based maintenance data, and eventually enable accurate decision-making (Brundage, M. P., Weiss, B. A., and Pellegrino, J., 2020). For example, in the railway maintenance business,

axle bearings are some of the most critical rolling stock components subject to strong safety constraints. In consequence, many conservative overhaul actions are scheduled preventively in the maintenance plan, which contains a lot of technical documentation about these mechanical assets. The completion of these actions, in turn, generates useful practical feedback on the shop floor following the inspection of the parts, which seeks degradation signals and compiles them in written maintenance sheets. Additionally, unexpected failures like grease leaks, hot axleboxes, or abnormal vibration records, get reported in an issue tracking system to be then fixed correctively. Considering all these environments together entails dealing with a large amount of text data that is oftentimes manually intractable, and NLP brings the automation potential to extract useful insights to advise the maintenance team, e.g., by identifying the most probable underlying root cause to a given problem. This approach is meant to increase the chances of success to fix the issue, minimize the risk of a recurrent failure, and thus maximize the availability of the fleet.

Interactive natural language interfaces help maintainers achieve a higher success rate and a lower task completion time, which lead to greatly improved user satisfaction (Su, Y., Awadallah, A. H., Wang, M., and White, R. H., 2018). However, many solutions require customization through the collaboration between data scientists and domain specialists, and each technical field poses its own challenges. In this sense, Technical Language Processing (TLP) presents a holistic, domain-driven approach, to use NLP in a technical engineering setting (Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S., 2021). In TLP, maintenance documents like work orders are relatively small in size and contain misspellings, domain-specific jargon, abbreviations, and non-standard sentence structure. Therefore, to tackle this particular context-dependent technical scenario, the field of causality is

Alexandre Trilla et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

<https://doi.org/10.36001/IJPHM.2022.v13i2.3127>

regarded as a direct description of what occurs when machines degrade, and the root-cause analysis becomes the means to obtain a reliable troubleshooting explanation for an abnormal failure. In fact, linguistic representation, such as the one found in TLP, is essentially a causal phenomenon (Stampe, D. W., 2008).

Causality is traditionally stratified into a three-layer hierarchy (Pearl, J., 2019): association (i.e., plain correlation or direction-free relationships), intervention (i.e., reasoning about the effects of actions), and counterfactuals (i.e., retrospective reasoning). In turn, Causal Inference (CI) aims to draw such detailed interpretations beyond mere associations from observational data using statistical tools to infer relational probabilities. CI distinguishes two broad classes of causal queries: forward causal questions or the estimation of “effects of causes”, and reverse causal inference or the search for “causes of effects” (Gelman, A., and Imbens, G., 2013). CI can also be conceptualized as a multitask learning problem with a set of shared layers among the factual and counterfactual outcomes (Alaa, A. M., Weisz, M., and van der Schaar, M., 2017). Similarly, decision-making is about predicting counterfactuals (Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M., 2017), and CI can potentially lead to more informed decisions (Zheng, M., Marsh, J. K., Nickerson, J. V., and Kleinberg, S., 2020). The difficulty here is that all these probabilistic quantities are not directly available in observational/factual data, so the CI problem needs to be converted into a domain adaptation problem to figure out the mechanisms that explain why observations occurred (Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A., 2020).

Understanding causality is considered as one of the current challenges for Machine Learning (ML) automation because ML models are ultimately driven by correlations in the data, and in general the causality implications of interest cannot be derived from them (Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, J., Schölkopf, B., Wüthrich, M., and Bauer, S., 2020). Therefore, counterfactual explanations are gaining prominence as a way to explain the decisions of a ML model (Barocas, S., Selbst, A. D., and Raghavan, M., 2019). The causality hierarchy, and the formal restrictions it entails, explains why ML systems can attain CI as long as they model the data beyond mere observed associations. Therefore, learning causal relations can be transformed into a supervised prediction problem once the data labels indicate the causal directionality, whether explicitly or implicitly (Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H., 2020; Shalit, U., Johansson, F. D., and Sontag, D., 2016). In this line of work, research in ML and language understanding have recently found a great deal of success using large neural networks, especially through Deep Learning (DL) (Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A., 2020; LeCun, Y. and Bengio, Y., and Hinton, G. E., 2015). These overparameterized and regularized models constitute one of the most important ideas in

the recent history of statistics, along with CI (Gelman, A., and Vehtari, A., 2020), and a straightforward way to learn causal effects and counterfactual outcomes with DL is to learn representations for features, i.e., to let the DL system automatically discover the most effective way to represent the data directly instead of hard-coding traditional language features. To this end, DL-based word embeddings may provide an interesting approach to represent linguistic causality (Li Y., and Yang T., 2018; Hancock, J. T., and Khoshgoftaar, T. M., 2020).

Specifically, Word Embeddings (WE) are dense, fixed-length word vectors, built using word co-occurrence statistics as per the distributional hypothesis (Almeida, F., and Xexéo, G., 2019). WE learn representations of high-level abstract concepts of the kind humans manipulate with language, away from the perceptual space, and they exhibit some geometric relational properties (Bengio, Y., 2017), which can ultimately be used to conduct lexical comparisons (Tan, L., Zhang, H., Clarke, C. L. A., and Smucker, M. D., 2015). Thus, this data representation can be regarded as an approach to cognition and artificial intelligence (Maguire, P., Mulhall, O., Maguire, R., and Taylor, J., 2015). Moreover, WE are computationally efficient (Levy, O., and Goldberg, Y., 2014), and therefore they need less data to successfully train statistical models (Goth, G., 2016), as is the case in TLP. Regarding semantics, WE also expose word senses (Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H., 2019), but they may experience the meaning conflation deficiency that arises from representing a word with all of its possible meanings as a single vector (Camacho-Collados, J., and Pilehvar, M. T., 2018). Nevertheless, WE constructed using arbitrarily contextualized language have further improved representational performance, possibly helping in the semantic disambiguation of machine decay (Levy & Goldberg, 2014; Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., 2018). In this line, WE also lead the way to process language in Prognostics and Health Management (PHM) because they display a high flexibility that is only attained by avoiding task-specific engineered features (Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020).

The troubleshooting objective pursued in this article is interesting for the PHM community to enhance the maintenance business (Leao, B. P., Fitzgibbon, K. T., Puttini, L. C., and de Melo, G. P. B., 2008). Realizing a comprehensive monitoring of system data, a timely detection of system abnormalities, and troubleshooting are all worthy goals, and the recent exponential growth of PHM patents is a point of support for these advantages (Liu, Z., Jia, Z., Vong, C.-M., Han, W., Yan, C., and Pecht, M., 2018). Current troubleshooting tools rely on fault tree analysis, extensive electronic manuals or expert system methods to assist the maintainer in identifying faulty system components (Naveed, A., Li, J., Saha, B., Saxena, A., and Vachtsevanos, G., 2012). The approach presented in this paper combines these complementary methods through the

exploitation of technical text data from different environments, which is aligned with the scope of PHM (Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S., 2021).

This work applies the CI paradigm to PHM using DL through a contextualized WE to better troubleshoot the root causes of failures and help improve their diagnostics. To do so, it exploits two different linguistic environments where causality is expected to be observed. On the one hand, an ontological reference framework based on a Failure Mode, Mechanism, and Effect Analysis (FMMEA), which provides a scholarly structure of causality driven by degradation. On the other hand, an actual record on Return On Experience (ROX), the data of which have been explicitly written for the purpose of explaining the root causes of the reported failures. In both environments, several experts inherently identify which properties of the observations describe spurious correlations unrelated to the causal explanation of interest, and which properties represent the phenomenon of interest, i.e., the stable invariant correlations (Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D., 2019). In this controlled analysis dealing with experimental data, invariant correlation implies causation. Therefore, DL and WE should be adequate tools to extract the textual regularities that represent causality (Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C., 2021), and thus they may be used to rate the level of agreement between CI theory and practice for troubleshooting. Specifically, a probabilistic Causality-Contextualized WE (CCWE) is trained with the ROX data, and the FMMEA-based failure ontology data is then used to evaluate the alignment between the two environments, which is expected to be reasonably high. This hypothesis is validated experimentally using the technical documentation related to rolling stock bogies. Figure 1 shows a diagram of the proposed analysis workflow for clarity.

The article is organized as follows: Section 2 describes the data, i.e., the bogie FMMEA and ROX records, the way the ontology has been created, and the strategy to build a CCWE. Section 3 conducts a graphical analysis of the whole failure network to discover structurally interesting points, a probabilistic analysis of the ROX-based CCWE to assess the causal relationships in practice, and the integration of the two perspectives, including a distributed representation of causality. Section 4 discusses the limitations of the proposed approach through the comparison with an alternative spectral embedding and the modeling of textual sequences. Finally, Section 5 concludes the manuscript showing how the concept of causality in bogie failures has been partially attained with the current technical documentation, and how it may be improved with the approach presented in this work.

2. MATERIALS AND METHODS

In this section, a FMMEA for bogies is used to build a failure ontology of their degradation, and a text database of ROX data

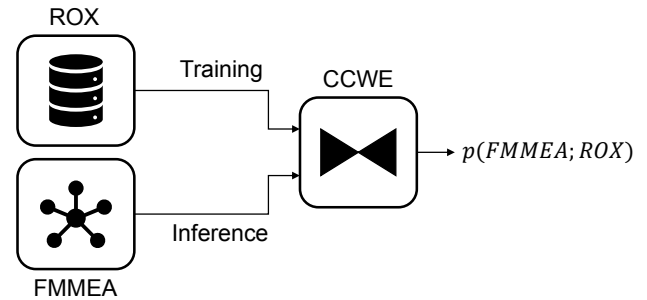


Figure 1. Diagram of the probabilistic analysis workflow performed in this work, which evaluates the level of agreement between two causality-rich environments: the Failure Mode, Mechanism, and Effect Analysis (FMMEA) on the theoretical side, and the Return On Experience (ROX) on the practical side. A Causality-Contextualized Word Embedding (CCWE) is developed to model and evaluate the relevant causal linguistic regularities.

is used to build a practical CCWE.

2.1. Failure Ontology

The FMMEA is an efficient tool to analyze system and component failures, and identify their main causes or mechanisms of failure (Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N., 2017). Knowledge of the failure mechanisms that are likely to produce the degradation that can lead to eventual failures in the monitored assets is important to succeed in the implementation of a PHM solution (Mathew, S., Das, D., Rossenberger, R., and Pecht, M., 2008). Therefore, the FMMEA is one of the tools used for the effective assessment of risk, and so it is a vital part of an organization's strategic management. However, it is costly to produce and hardly reusable due to its text-based description in natural language (Ebrahimipour, V., Rezaie, K., and Shokravi, S., 2010). To overcome this situation, an ontology-based solution is advised to extract and reuse FMMEA knowledge from the available text documents (Rehman, Z., and Kifor, C. V., 2016).

An ontology is a network of standard concepts and terms in a given domain that shows their properties and the relations between them to represent knowledge (Ebrahimipour, V., Rezaie, K., and Shokravi, S., 2010). There is a growing interest in the potential value of ontologies to codify structures of meaning for maintenance (Sexton, T., Hodkiewicz, M., Brundage, M. P., and Smoker, T., 2018). To this end, TLP is the way to go to automatically extract valuable insights regarding the many facets of reliability, maintenance, and planning (Navinchandran, M., Sharp, M. E., Brundage, M. P., and Sexton, T. B., 2019). The ontology augments human decision-making by relying on diversified information (Polenghi, A., Roda, I., Macchi, M., and Pozzetti, A., 2022), especially when real-life maintenance data is used in its design. Conforming to the vocabulary that is widely used by maintenance professionals and practitioners is

a major catalyst for widespread acceptance and uptake (Karray, M. H., Ameri, F., Hodkiewicz, M., and Louge, T., 2019). Additionally, to tackle CI with the ontology, its topology needs to represent a Structural Causal Model (SCM) framework because its organization is essential for performing causality learning tasks such as counterfactual reasoning (Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y., 2021). Explicitly, a SCM consists of a set of explanatory variables, outcome variables, and unobserved variables, connected by a set of functions that determine their relational values (Pearl, J., 2009).

For the analysis of bogie failures framed in this work, a FM-MEA approach is recommended to reduce blindness, subjectivity, and over-reliance on the personal experience (Li, Y.-H., Wang, Y.-D., and Zhao, W.-Z., 2009). And for the successful application of CI, assumptions about the mechanisms underlying the observed data also need to be specified (Sharma, A., and Kiciman, E., 2020). To this end, the approach proposed by Atamuradov and colleagues is taken for reference in this work, and thus its contents are not questioned here (Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N., 2017). Their failure analysis defines three fields that are described as follows, along with the related causal structure:

Failure Mechanism Fundamental manner in which a component can fail → Unobserved variable that is the Root Cause of an observed Problem, e.g., fatigue or wear

Failure Mode Manner by which a failure is physically observed, although in certain contexts, the Failure Effect (i.e., the impact of the Mechanism) can also be found in this field → Outcome variable that represents a Problem that is experienced, e.g., surface defects, rotation difficulty, or reduction of suspension effect

Component Explanatory variable that describes the context of a Problem, e.g., wheel or gearbox

Components are related to Failure Modes, which in turn are then related to Failure Mechanisms. If these relationships are likened to an ISO 13379 standard causal tree with faults, symptoms, and descriptors (ISO, 2003), the resulting failure ontology is shown in Figure 2, where the directed edges indicate the (assumed) direction of causation (Imbens, G. W., 2020).

2.2. Return On Experience

ROX is a holistic approach to understand and increase the value of investments across customer, employee, and leadership experience (PwC, 2019). It is strictly related to the First Time Right management principle, which aims to minimize the number of product issues that get past design release and cause rework, leading to dissatisfied customers (Leuenerger, H., Puchkov, M., and Schneider, B., 2013). Specifically, ROX is a data-driven quality strategy that focuses on identifying and eliminating the root cause of the problems and ensure

that the improvement is sustained (Smetkowska, M., and Mru-galska, B., 2018). To this end, tagging and curating already existing textual data can be a first step toward structuring content (Sexton, T. B., and Brundage, M. P., 2019), but this work goes beyond this step and processes data that have been specifically written for the purpose of describing the causal sources of the reported problems. Therefore, unlike regular observational data, ROX records are hardly marred by selection biases, confounding factors, and other such weak points, and thus they may be treated as experimental or interventional data.

The ROX database of use in this work contains around 500 records written by many experts following a feasible collaborative approach (Hastings, E. M., Sexton, T., Brundage, M. P., and Hodkiewicz, M., 2019). However, different technicians rarely describe the same Problem in an identical manner or register (Conrad, S., 2019). This leads to description inconsistencies within the database and makes it difficult to categorize issues or learn from similar causal relationships (Sharp, M. E., Sexton, T. B., and Brundage, M. P., 2017). Therefore, a statistics-based TLP approach is needed to put the focus on factual data and strip grammatical artifacts, e.g., by filtering out stop words, lemmatizing, etc. This provides a systematic methodology to create computable knowledge (Sexton, T., Hodkiewicz, M., Brundage, M. P., and Smoker, T., 2018).

By definition, plain text data are intrinsically unstructured. However, in the ROX database each record conducts a specific troubleshooting analysis in isolation, and the causal connections are organized into the following fields:

Problem Subject title, description of the reported Failure Mode, and details of its technical impact.

Root Cause Description of the Failure Mechanism of the issue following an investigation, and the main reason of non-detection.

Business context Strategic unit: trains, rail services, rail control, and infrastructure.

System context Technical scope: air supply, passengers, roof, door, and bogie.

Issue context Domain category: mechanical, documentation, electrical, and assembly.

Table 1 shows some examples of bogie ROX database entries to illustrate the nature of these data (note that the majority of the instances are mechanical issues).

To further understand the characteristics of these technical text data, which justifies the TLP-based approach, Figure 3 shows the power-law distribution of its ranked word frequencies compared to what is expected in natural language (Zanette, D. H., and Montemurro, M. A., 2005). Note that the technical language curve has a positive offset with respect to natural language. This increased word frequency spectrum may indicate that this technical language shows a reduced vocabulary and

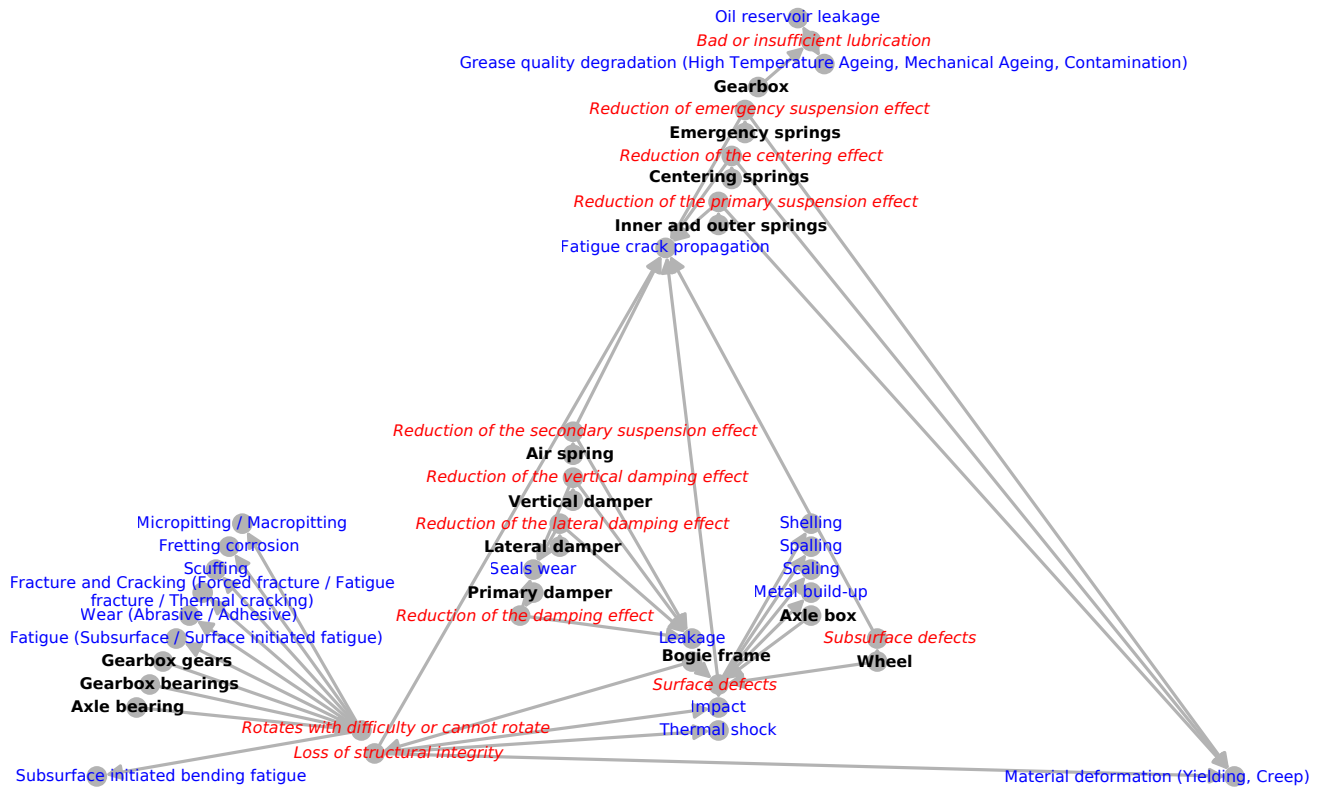


Figure 2. FMMEA-based bogie failure ontology linking Component (black boldface) to Failure Mode (red italics) and then to Failure Mechanism (blue).

therefore the same words may need to be used more often. In a similar descriptive vein, Figure 4 shows the distribution of technical ROX text lengths as word counts per record along with some comparative hints regarding natural language. Note that the statistical ROX length mode is around 8 words, which is far from the optimum contemporary readability indication of 17 words (DuBay, W. H., 2004). Such short texts have some unique characteristics that make them difficult to handle. For instance, they do not always observe the syntax of written language, they contain limited context, and they give rise to ambiguity as more than one meaning may be conveyed, leading to vagueness and confusion (Wang, Z., and Wang, H., 2016). Moreover, the ROX length distribution shows a tail of longer texts that get increasingly difficult to read, and also over 35 words the quality of a language model decreases rapidly (Bahdanau, D., Cho, K., and Bengio, Y., 2015).

2.3. Causality-Contextualized Word Embedding

The original conception of a WE related a single word to its local context given a shallow window of proximity (Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013). However, this principle does not hold for CI because the context of the

related texts is different. In this work, the goal is to learn the causal relationships between Problems (i.e., Failure Modes) and their Root Causes (i.e., Failure Mechanisms) through their respective textual expressions. To do so, a binary-valued Bag-Of-Words (BOW) model is considered to account for the presence of multiple words concurrently (Le, Q., and Mikolov, T., 2014). Note that the syntax is not retained as this model focuses on the overall semantics through the lexicon. In turn, the input and output vocabularies are also dependent on their causal roles, and regarding that an effective method depends on the size of the vocabulary (Chen, W., Grangier, D., and Auli, M., 2015), both Root Cause and Problem lexicons are considered in the present WE model.

The proposed implementation of the CCWE for troubleshooting is based on an encoder-decoder DL architecture using the causal concept of refinements (Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C., 2021), see Figure 5. Root Causes are probabilistically modeled given their Problems and some Context, which is a situational hint to enhance language models (Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., and Jiang, M., 2020), and may be stripped from the model once trained. The CCWE is exploited with a contrastive esti-

Table 1. ROX database examples of bogie system failures reported by maintenance services.

Issue context	Problem	Root Cause
mechanical	vertical damper failure, sealing defect	as per supplier investigation report the failure mode is the primer glue departed from the metal parts. considers it was because the metal parts were not cleaned well while in the pre treatment process the primer glue can't adhere well to the metal parts so it will cause debonding issue during operation.
mechanical	anti roll bar assembly knocking noise, excessive noise	a light stick slip phenomena is the root causes of the noise. it is decided to change the knuckle as per updated design from supplier hyed for one complete train set. currently in claim situation with supplier for them parts are compliant to specification.
mechanical	oil leakage from gear box unit, loss of tightness	as per supplier rca it is confirmed that the gear lubricating oil from the drainage hole leakage caused by the labyrinth ring tw of roundness error. oil leakage causes in the process of the part in ngc in sheet2 the process of operation not suitable for the mode of transportation easy to cause roundness error of deformation when parts fall off or pressure deformation.
assembly	conical spring bonding issue, loss of regulation	debonding beetwen rubber material and steel frame incorrect handling of adhesived parts by operators before putting them into the mould. the cleanliness of localized area is jeopardized and it disturb the bonding process between rubber and interface. it was not possible to detect during the validation tests the parts tested did not presented failure. the issue happens when submitted to load sometimes with few milage or more than 150.000 km for example.

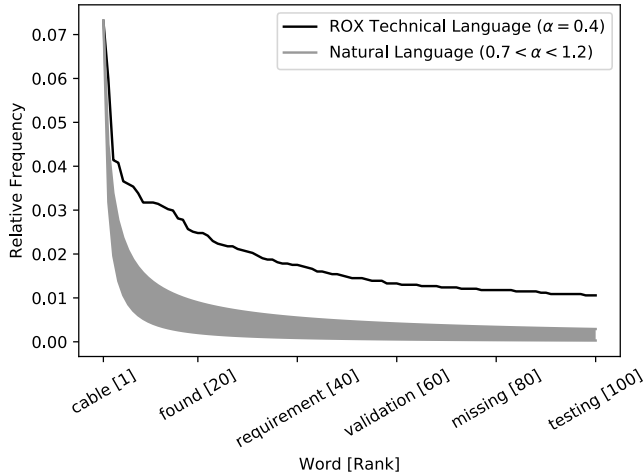


Figure 3. Ranked word relative frequency distribution of technical ROX text data versus natural language. The exponent of the power laws is shown in brackets.

mation framework, which discriminates between the observed data and some artificially generated noise (Gutmann, M., and Hyvärinen, A., 2010; Mnih, A., and Teh, Y. W., 2012; Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., 2013). This approach is attained through jointly learning a series of nonlinear logistic regressions using an output logistic activation function and a cross-entropy cost criterion for training. Bias terms are also considered because of the multiple-word instances with different lengths (there is no basis to assume that the embedding will be centered around the origin). Also, being a DL solution the model is expected to be overparameterized, so the use of Dropout layers is recommended to manage words that belong to regions of poor overlap in the feature space (Alaa, A. M., Weisz, M., and van der Schaar, M., 2017). Specifically, the input layer is followed by a Dropout layer

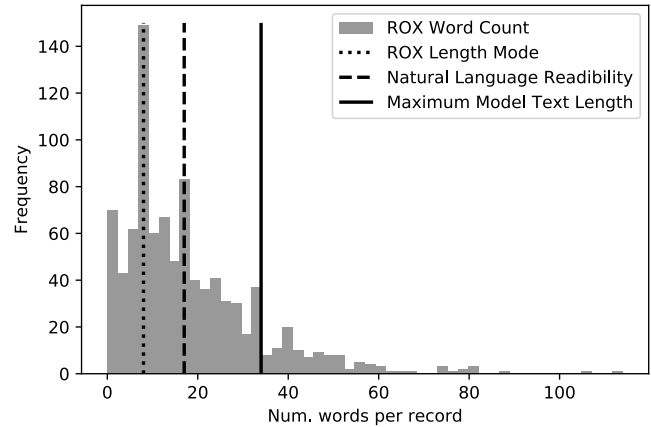


Figure 4. Word count frequency distribution of technical ROX text data records along with natural language readability indications.

to deal with long texts because these are more likely to have words deactivated, therefore equaling their potential impact to that of shorter instances. And the embedding layer, which is smaller than the BOW-based layers, is also followed by another Dropout layer to adjust its representational expressiveness and manage ambiguity more effectively (Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H., 2019).

The proposed CCWE model gives the following probability directly:

$$p(\text{Root Cause} | \text{Problem}, \text{Context})$$

However, an explicit formulation through the embedding bottleneck layer is advantageous to study the geometric properties of its distributed representation, see Eq. (1).

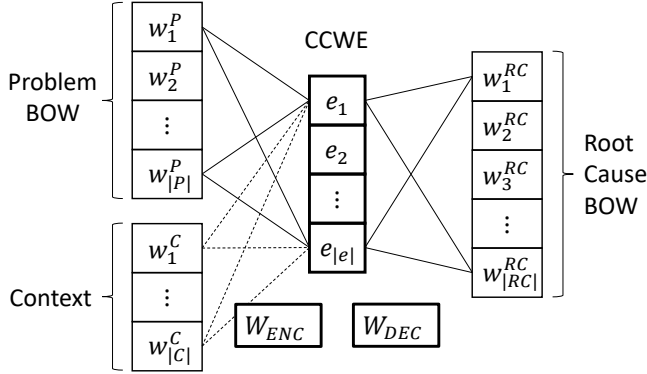


Figure 5. Encoder-decoder DL architecture of the CCWE in inference mode. Dropout layers are used in training mode only, and are thus not shown here for clarity.

$$\begin{aligned} CCWE &= W_{ENC} \cdot (Problem, Context) \quad (forward) \\ &\sim W_{DEC}^+ \cdot \text{logit}([Root Cause]) \quad (backward) \end{aligned} \quad (1)$$

Note that the backward equation requires the inversion of the non-square decoder matrix W_{DEC} , which is not possible. In this case, a least-squares approximation is used through its pseudoinverse W_{DEC}^+ . Also note that the *logit* function cannot be applied to a binary-valued BOW vector because it leads to an asymptotic overflow. In this case, the values of the $[Root Cause]$ vector are clipped to 0.2 (false) and 0.8 (true). These bounds are driven by the extrema of the second derivative of the logistic function and prevent its saturation.

Finally, the distributed representation of causality is to be exploited through the Principal Components (PC) of the CCWE and the cosine distance between Root Cause and Problem BOW vectors (Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., 2013). The angle they form in the PC space is a common textual similarity metric utilized in semantic classification and search (Tan, S., Zhou, Z., Xu, Z., and Li, P., 2019). And taking into account that the cosine similarity becomes less predictive as the dimensionality increases (Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H., 2019), the PC representation is typically reduced to two dimensions following the customary practice in NLP research.

3. RESULTS

Causal prediction is not a typical downstream NLP task apt for evaluation. Therefore, the experiments conducted in this section have been compared with human judgments on word relations, i.e., an intrinsic evaluation (Bakarov, A., 2018). Explanations have been provided through graphs, feature importance (e.g., word probabilities), visualizations (e.g., spectral analysis), and concrete examples (Mothilal, R. K., Sharma, A.,

and Tan, C., 2020).

3.1. Causal Graphs

Graphs are a powerful representation formalism that can be applied to a variety of aspects related to language processing (Mihalcea, R., and Radev, D., 2011). With a proper choice of nodes and edge drawing criteria and weighing, graphs can be extremely useful for revealing regularities and patterns in the data (Nastase, V., Mihalcea, R., and Radev, D., 2015). Additionally, causal graphs reduce the adverse impact of latent variables or noise (Bahadori, M. T., and Heckerman, D. E., 2021). This section studies the failure ontology as a causal graph to detect confounders (i.e., common root causes) as forks, and colliders (i.e., common problems) as inverted forks. To get an overview of these characteristics, centrality measures have been used to pinpoint the most important nodes of the resulting graphs.

On the one hand, the degree centrality $C_D(v)$ states that the important nodes v are the ones that have many connections (Mihalcea, R., and Radev, D., 2011), see Eq. (2), where V is the total number of nodes in the graph, and d is the distance between two nodes, i.e., the minimum number of vertices that separate them. The application of this criterion is shown in Table 2 as a ranking of nodes, and Figure 6 shows a graph that preserves the ontological relationships driven by this ordered arrangement. According to the degree centrality indicator, the confounders are the nodes related to the Failure Modes of the suspension components (i.e., springs, damper...), and the colliders are its Failure Mechanisms (i.e., fatigue crack, material deformation, leakage, and the wear of seals).

$$C_D(v) = \frac{1}{V} \sum_{\forall v' \neq v} x, \text{ where } x = \begin{cases} 1 & \text{if } d(v', v) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

On the other hand, the closeness centrality $C_C(v)$ states that the important nodes v are the ones that are near other nodes v' (Mihalcea, R., and Radev, D., 2011). This proximity indicator is calculated as the inverse of the sum of the path lengths from a given node to all the other nodes, see Eq. (3). The application of this criterion is shown in Table 3 as a ranking, and Figure 7 shows the corresponding graph that preserves the ontological relationships. According to the closeness centrality indicator, the confounders are the nodes related to the Failure Modes of the bearings: surface defects and rotation difficulty. In general, note that the nodes with the greatest centrality measures are not densely connected among themselves (some even show few connections), thus there are many peripheral items to be taken into consideration.

$$C_C(v) = \frac{V-1}{\sum_{\forall v' \neq v} d(v', v)} \quad (3)$$

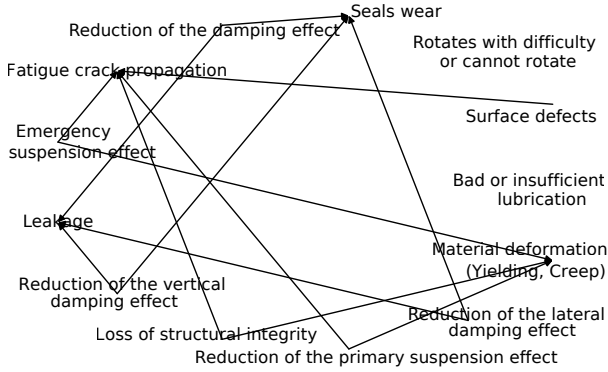


Figure 6. Failure ontology subgraph driven by the nodes with the greatest degree centrality.

Table 2. Ranking of failure ontology nodes according to their degree centrality score.

Failure Ontology Node	Degree Centrality
Rotates with difficulty or cannot rotate	0.2273
Surface defects	0.2045
Fatigue crack propagation	0.1591
Loss of structural integrity	0.1136
Leakage	0.0909
Material deformation (Yielding, Creep)	0.0909
Seals wear	0.0682

Table 3. Ranking of failure ontology nodes according to their closeness centrality score.

Failure Ontology Node	Closeness Centrality
Fatigue crack propagation	0.2121
Leakage	0.1212
Material deformation (Yielding, Creep)	0.1212
Seals wear	0.0909
Impact	0.0710
Surface defects	0.0682
Rotates with difficulty or cannot rotate	0.0682

3.2. Causal Lexical Probabilities

This section conducts a preliminary study of the sensitivity of the CCWE built with the bogie ROX data. The dimensionality of the BOW for the Problem is $|P| = 1591$, for the Root Cause it is $|RC| = 2210$, and for the embedding it is $|e| = 300$. This configuration yields a model with more than 1M trainable parameters. This WE has been trained using cross-validation with a train/test data split of 80%/20%, and the resulting binary accuracy is 0.9894. This learning result indicates that the memorized word relationships of the CCWE are likely to provide reliable causal associations for ROX. To illustrate the troubleshooting capacity of the CCWE, Table 4 shows some

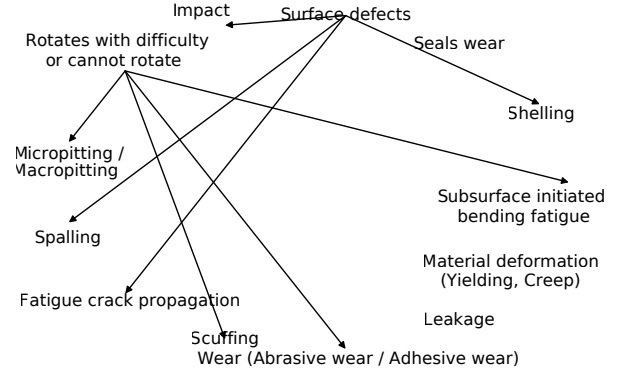


Figure 7. Failure ontology subgraph driven by the nodes with the greatest closeness centrality.

Table 4. Generic troubleshooting word examples obtained with the CCWE.

Problem	Possible Root Cause (Probability)
oil leak	attached (0.8849), measured (0.6137), hole (0.5733), pressure (0.2372)
bearing	tightening (0.0659), vibration (0.0639), shock (0.0495), assembly (0.0394)
gear box	design (0.9062), tolerance (0.9061), oil (0.8703), pressure (0.8237)

generic word examples.

In general, the Root Cause outcomes of the CCWE with high probability are reasonable words that belong to the same semantic field of the given Problems. Note that the probabilities for the “bearing” component are an order of magnitude lower than those for “oil leak” and “gear box”. This result may be due to the specificity of causal words like “tightening”, compared to common words like “attached” or “design”. However, there are also some noise words that typically appear in the BOW of the Root Cause, such as “please”, “report”, “reference”, “part”, etc. This is attributed to the way the experts provide standard ROX feedback. Also, the arrays of output probabilities are mostly comprised of low values, and this is mainly explained by the large space of BOW dimensionality, which leads ROX instances to be sparse.

The geometrical characteristics of the obtained linguistic distributed representation are shown in Figure 8. Note that to obtain this rendering, both the forward encoder and backward decoder equations of the CCWE are needed. This distribution shows that the Root Causes are concentrated in the center, whereas the Problems are spread across the PC space. Thus, the cosine similarity metric is needed to align them within the α angle, yielding a circular sector of causal likelihood. A detailed example of the alignment between a generic Problem like “noise” and its potential Root Causes is shown in Figure 9. The results illustrate the incertitude of the derived causal rep-

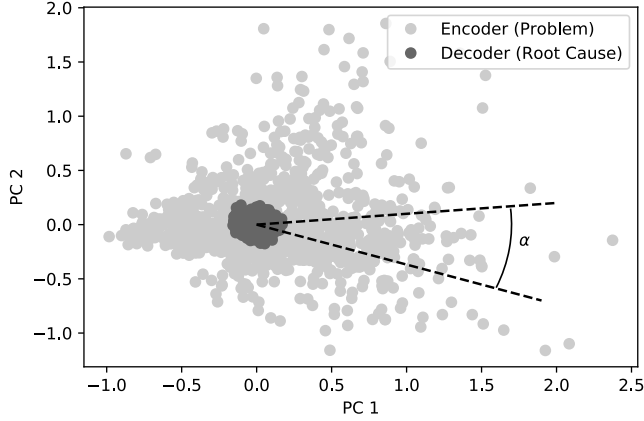


Figure 8. PC of the CCWE activations and the application of the cosine distance similarity measure showing a circular sector α of causal likelihood.

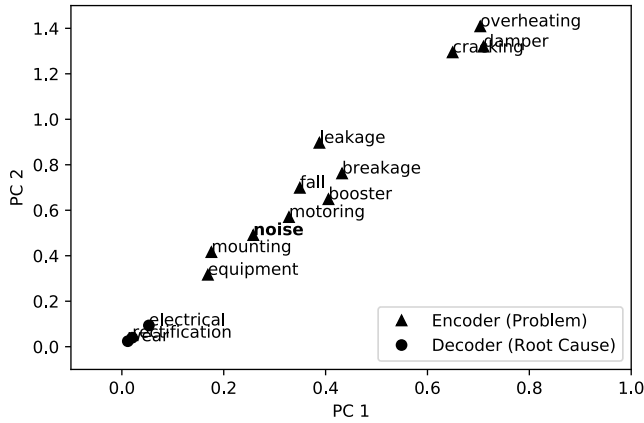


Figure 9. Detailed example of a cosine distance lower than five degrees ($\alpha < 5^\circ$) between the generic “noise” Problem, its nearest Root Causes, and other close/similar Problems.

resentation as the nearest Failure Mechanisms are “electrical/rectification” and “wear”. In addition, many reasonably related “noise” Problems (sharing the same Root Causes) are also shown, e.g., “motoring”, “breakage”, “leakage”, “cracking”, etc.

3.3. Troubleshooting Integration

This section determines if the relationships in the FMMEA-based failure ontology correspond to high ROX-based causal probabilities. To do so, the evaluation of whole Failure Mode texts (as Problems P) is conducted by taking the average probability $\bar{p}_{ROX}(RC|P)$ of the Root Cause RC words appearing in the reported Failure Mechanisms, see Eq. (4), where N represents the words in the text being evaluated. Table 5 shows the top-ranking failures that have been obtained. These results indicate that the leading issues are related to springs and wheels, which the latter is in accord to previous knowledge (Trilla,

A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X., 2021). Also note that they are mostly linked to the main confounders, i.e., the common root causes, of the failure ontology.

$$\bar{p}_{ROX}(RC_i|P_i) = \frac{1}{N} \sum_{w \in N} p_{ROX}(RC_i^w|P_i) \quad (4)$$

$i \in \text{FMMEA Failure Ontology}$

In addition to this direct FMMEA/ROX relationship, it is also necessary to determine if the cross-failure probabilities are low and thus assert that the proposed approach shows a discriminative property. This alignment study has been determined using the Cross-Probability Difference (XPD) variable, defined by Eq. (5) as the difference between the direct causal probability and the anti-causal probabilities. Note that positive probability differences represent a good alignment between Failure Mode and Mechanism i , whereas negative differences mean that other Failure Mechanisms j are more relevant (according to ROX) than the one stated in the FMMEA-based failure ontology.

$$XPD(i) = \bar{p}_{ROX}(RC_i|P_i) - \bar{p}_{ROX}(RC_j|P_i) \quad (5)$$

$\forall j \neq i$
 $i, j \in \text{FMMEA Failure Ontology}$

Regarding the distribution of the XPD variable, see Figure 10, the majority of the FMMEA statements are aligned (71.32% of strictly positive values). The clearest textual expressions are driven by the centering springs component. Nevertheless, there are many cases where the difference is too small to extract strong conclusions, as is shown by the high peak around 0. Maybe this is due to averages including missing terms, e.g., specific Failure Mechanism words like “spalling”, “scaling”, “scuffing”, and “pitting” do not appear in ROX. In addition, there are some outlier instances showing a large misalignment, i.e., $XPD < -0.06$. Some examples are listed as follows:

- Bogie frame, Surface defects → Material deformation (Yielding, Creep)
- Bogie frame, Surface defects → Shelling
- Wheel, Surface defects → Material deformation (Yielding, Creep)
- Vertical damper, Reduction of the vertical damping effect → Metal build-up

All these results may be taken for different signs of poor writing, and thus may also be an indication to rephrase those statements and improve the meaning they convey.

To conclude the integration analysis, Table 6 shows an indirect evaluation of the application of the ROX-based causality to the FMMEA-based failure ontology through the cosine distance as the PC vector angle similarity. Bearings and suspension components populate this ranking, which is quite similar to

Table 5. Ranking of ROX-based average probabilities driven by FMMEA failure ontology.

Component	Failure Mode	Failure Mechanism	\bar{P}_{ROX}
Centering springs	Reduction of the centering effect	Material deformation (Yielding, Creep)	0.0308
Wheel	Surface defects	Shelling	0.0208
Emergency springs	Reduction of emergency suspension effect	Material deformation (Yielding, Creep)	0.0184
Inner and outer springs	Reduction of the primary suspension effect	Material deformation (Yielding, Creep)	0.0147
Bogie frame	Loss of structural integrity	Material deformation (Yielding, Creep)	0.0071
Bogie frame	Loss of structural integrity	Fatigue crack propagation	0.0021
Bogie frame	Loss of structural integrity	Impact	0.0019

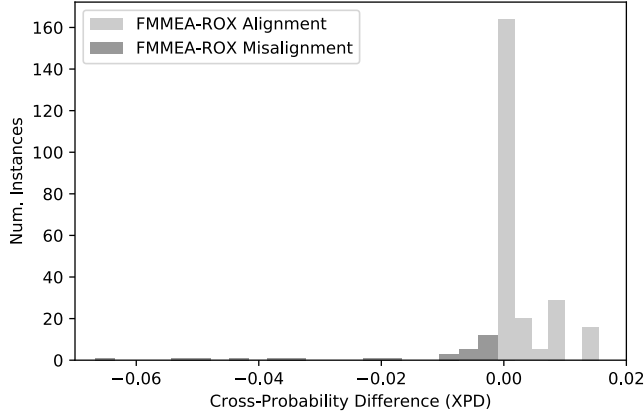


Figure 10. Cross-Probability Difference (XPD) distribution visualized through the histogram.

the one driven by the causal probabilities (a slight reordering is observed, though). In fact, angles and probabilities score a Pearson correlation coefficient of -0.65 , so the previous probability-driven conclusions are likely to be largely extrapolated in this causal distributed representation. Therefore, the FMMEA entries that display wide ROX angles may indicate that a rephrasing would be beneficial to increase the comprehension of their text (Ansari, F., 2020). Anyhow, all these results show that the FMMEA ontology relations can be reasonably weighted either via ROX causal probability or distance scores, and thus obtain a SCM to validate the CI approach using a DL-based contextualized WE.

4. DISCUSSION

Up to this point, after having completed the workflow procedure, the discrepancy between FMMEA and ROX has been solely attributed to lexical imprecision between the same causality principles expressed in a particular environment, context, or perspective. However, there may be other sources of epistemic uncertainty that could help explain this divergence. This section addresses some particularities about the proposed CCWE model.

For example, by the Independent Causal Mechanisms principle, the causal generative process of a system's variables is composed of autonomous modules that do not inform or

influence each other (Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y., 2021). In the troubleshooting probabilistic case tackled in this work, this would imply that the conditional distribution of each Root Cause variable (i.e., the Failure Mechanism) given its Problem (i.e., its Failure Mode) did not inform or influence the other causes. The presented CCWE does not respect this principle because of its multilayer neural topology trained using the standard backpropagation procedure: the encoder layer is influenced by all of the output cause variables, and this, in turn, affects all the predictions through the forward propagation. However, this could also be seen as an advantage from a multitask learning perspective (Crawshaw, M., 2020).

Additionally, performance gains of word embeddings are due to certain system design choices such as dynamically sized context windows and hyperparameter optimizations, rather than the embedding algorithms themselves (Levy, O., Goldberg, Y., and Dagan, I., 2015). This argument leaves the door open to considering *chance* as the ultimate explanatory factor for the results obtained. At the same time, it motivates further research study on DL-based WE.

4.1. Spectral Embedding

Probabilistic models like the CCWE can be viewed as directed graphical models (Salakhutdinov, R., and Hinton, G., 2009). As such, their learned knowledge may be interpreted using a graph spectral embedding or clustering technique. A suitable approach to extract this representation is through the factorization of the Laplacian matrix $L = D - A$, which is a measure of the local derivative of the graph (Mihalcea, R., and Radev, D., 2011). D represents the degree matrix (i.e., the amount of node incoming or outgoing links), and A represents the adjacency matrix (i.e., the causal word relations). After extracting the eigencomponents of L , similar nodes must have embeddings that are close to one another (Cai, H., Zheng, V. H., and Chang, K. C.-C., 2018), and thus the Euclidean distance could be adequate for the similarity comparisons. This section explores this proximity property in the present causal degradation environment.

Figure 11 shows a representation of the two largest Laplacian eigenvectors, which that aim to capture the maximum information (in the form of variance dispersion) of the em-

Table 6. Ranking of ROX-based cosine distance (angle similarity) driven by FMMEA failure ontology.

Component	Failure Mode	Failure Mechanism	α
Gearbox bearings	Rotates with difficulty or cannot rotate	Wear (Abrasive wear / Adhesive wear)	0.5995
Vertical damper	Reduction of the vertical damping effect	Seals wear	5.9321
Axle bearing	Rotates with difficulty or cannot rotate	Wear (Abrasive wear / Adhesive wear)	14.6010
Primary damper	Reduction of the damping effect	Seals wear	15.1803
Centering springs	Reduction of the centering effect	Material deformation (Yielding, Creep)	17.9046
Inner and outer springs	Reduction of the primary suspension effect	Material deformation (Yielding, Creep)	20.4794
Wheel	Surface defects	Metal build-up	23.2299

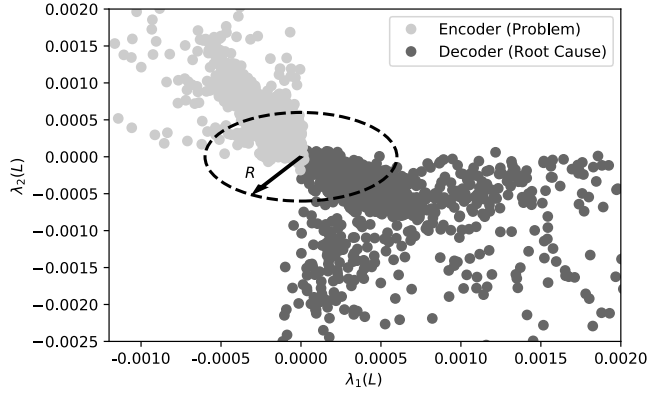


Figure 11. Largest Laplacian eigenvectors λ of the CCWE directed graph and the application of the Euclidean distance similarity measure showing a circle of causal likelihood R .

bedded causal data. Given the directed bipartite structure of the troubleshooting scenario tackled in this work, where the same word can be used to describe both the Root Cause and the Problem, two degree matrices have been used: one with the Problem word nodes (output degrees only), and the other with the Root Cause word nodes (input degrees only). Finally, their representations have been overlapped, showing that the cause/effect separation is preserved in this low-dimensional illustration. However, only the central region where the two causal roles meet seems to be amenable to any further inference assessment.

Figure 12 shows a more detailed example over the generic “pressure” Problem. All the Failure Mechanism terms that appear seem reasonable given this Failure Mode, e.g., “loop”, “zero”, “leak”, etc. However, in this case, the associated probabilities seem to be unrelated to the distance scores. Moreover, trying to replicate the “noise” Problem used before results in incomprehensible results due to the vast amount of terms that rapidly appear as the radius R is increased. Maybe the factorization of the Laplacian matrix, which is strictly defined for an undirected graph, built over a directed graph is flawed and needs further attention.

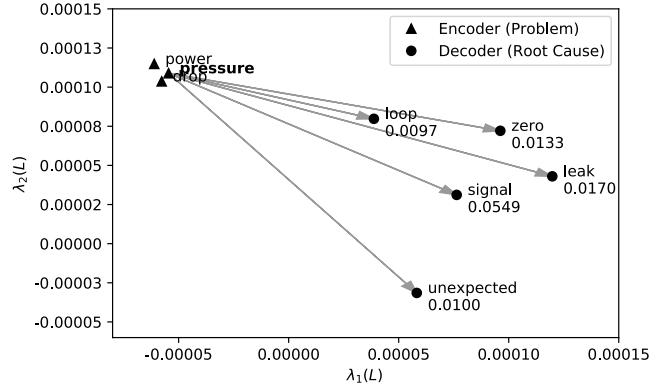


Figure 12. Detailed example of spectral embedding over the generic “pressure” Problem. In this troubleshooting scenario, arrows point toward the potential Root Causes, and the related probabilities are also shown under the words.

4.2. Language Modeling

In previous sections it has been shown that the lexicon per se is sufficient to produce reasonable causal probabilities. However, the principle of semantic composition states that the meaning of a phrase can be derived from the meaning of the words that it contains as well as the syntax that binds them (Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., and Daumé III, H., 2014). Likewise, a WE captures syntactic and semantic regularities (Mikolov, T., Yih, W.-t., and Zweig, G., 2013). Consequently, a WE could be able to compose meaningful phrases and thus build a language model.

Language models learn linguistic knowledge, store relational knowledge present in the training data, and may be able to answer structured queries (Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S., 2019). To do so, neural encoder-decoder models pioneered by machine translation were proposed to achieve the goal of mapping input text to output text (Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y., 2014). An encoder network first reads and represents a source sentence into a fixed-length vector, and a decoder network then outputs a target sentence from this encoded vector. This encoder/decoder architecture can also be extended to deal with corpora and vocabulary sizes, and complex, long term

structures of language (Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y., 2016). Eventually, encoder and decoder are jointly trained to maximize the conditional probability of a correct relationship, which is conceptually equivalent to what is pursued in the WE but this time considering the sequentiality of words as an additional embedded context (Liu, Q., Kusner, M. J., and Blunsom, P., 2020). This heteroassociative property is explored in this section to relate Root Causes to Problems for long texts.

The specific implementation adopted in this work is based on the Sequence-to-Sequence (S2S) approach. S2S applies recurrent neural networks to problems whose input and output sequences have different lengths with complicated and non-monotonic relationships (Sutskever, I., Vinyals, O., and Le, Q. V., 2014). Specifically, standard Long Short-Term Memory (LSTM) networks are used due to their superior performance for small corpora, as is the case in TLP, instead of more popular models based on Transformers (Ezen-Can, A., 2020). Also, model awareness of the context (e.g., through the WE) helps understand the semantic meaning of an input sequence and generate a more informative response (Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., and Jiang, M., 2020). Considering all these points, Figure 13 shows the diagram of the proposed causality-contextualized S2S language model using the LSTM and the CCWE. Note that given the sequential nature of S2S, the input/output interface to the system is no longer a BOW but a one-hot encoded single-word vector, i.e., words are presented and retrieved from the language model on a one-by-one basis.

Table 7 shows the plain Root Cause outputs obtained from the system given potential generic Problems. In light of these results, the causality-contextualized language model exhibits the performance of a “pidgin”, and this is mainly attributed to the strict lexicon-driven text preprocessing stage. The model does not retrieve the ROX entries literally. Instead, it displays a generalization capacity using vague words (e.g., most Problems are blamed on “reporting” as their Root Cause). Such pathological utterances, also known as *hallucinations*, are common with S2S (Lee, K., Firat, O., Agarwal, A., Fannjiang, C., and Sussillo, D., 2018). And due to the discrepancy between this vaguely generated text and the detailed ROX reports, the exposure bias problem that usually affects such autoregressive language models is increasingly more penalizing for technical language (Wang, C., and Sennrich, R., 2020). Also, input Problems need to be provided using long, elaborate and verbose descriptions, otherwise the model outputs nothing (i.e., long chains of padding symbols). This may be attributed to the most critical components of the LSTM cell, i.e., the forget gate and the activation function (Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J., 2017).

Finally, there are diminishing returns with increasing the scale of model parameters, dataset size, and training computation, because these variables are power laws (Kaplan, J., McCann-

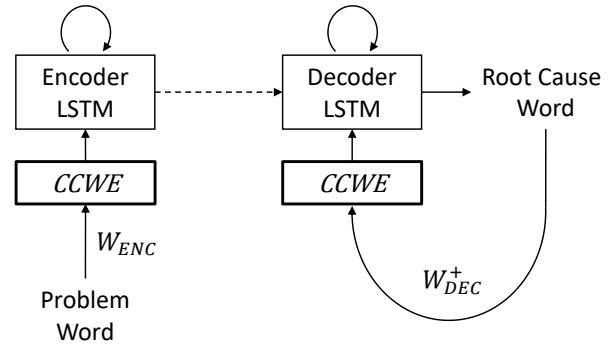


Figure 13. Diagram of the causality-contextualized S2S language model using the LSTM and the CCWE.

Table 7. Plain troubleshooting Root Cause sentences generated by the causality-contextualized language model given potential generic Problems.

Problem	Root Cause
oil leak found on bogie, gear box, and wheel at high speed	report design
hot axle box bearing	assembly
traction motor caught fire, smoke alert on commercial service	report inspection
noisy blower does not turn: power electronics are not available	report failure part

dlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D., 2020), so the potential for significant improvement needs to be driven by a complementary source of knowledge, such as the FMMEA, as it has been researched in this work. The causal structure of use here shows 19 Failure Mechanisms for 12 Failure Modes regarding 14 components, so further refinements (or generalizations) may be observed if these values are augmented.

5. CONCLUSION

This work describes a first exploratory work on how the Causal Inference paradigm may be applied to troubleshooting rolling stock bogies through the extraction of linguistic knowledge from FMMEA and ROX text data using graphs and contextualized word embeddings. The overall conclusions indicate that the inference of causality has already been attained with the available theoretical and practical documentation, showing a consensus greater than 70%. Interestingly, though, some disagreement between Root Cause and Problem has arisen in a few areas, leading to poor diagnosis results, and potentially indicating that textual expression improvements are necessary in the technical materials.

The central piece of this research is the construction of a neural word embedding that differs from the state of the art, which is focused on modeling the local context of a single

word. The proposed model jointly embeds two whole textual instances that belong to different (causal) contexts. In terms of evaluation, given that CI is not a well-defined task in language processing, the results may be questioned due to their strict dependence on subjective human criteria. This is a clear point of general improvement (beyond the specific purposes of this work) toward the fair assessment of other related CI approaches such as the Twin Networks method to estimate the probabilities of causation (Vlontzos, A., Kainz, B., and Gilligan-Lee, C. M., 2021), the causal regularization of neural networks to improve their interpretability (Bahadori, M. T., Chalupka, K., Choi, E., Chen, R., Stewart, W. F., and Sun, J., 2017; Shen, Z., Cui, P., Kuang, K., Li, B., and Chen, P., 2018), or the learning of causally disentangled representations using Variational Autoencoders (Suter, R., Miladinović, D., Schölkopf, B., and Bauer, S., 2019; Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J., 2020).

In terms of application, a direct implementation of this developing approach could be driven by a retrieval-augmented generation system for work orders to advise the maintenance team by identifying the most probable underlying root cause to a given problem, and reduce both the time to action and asset downtime while increasing the safety of the railway service (Ansaldi, S. M., Agnello, P., Pirone, A., and Vallerotonda, M. R., 2021). This enhanced troubleshooting system would equip a model that combines pre-trained parametric memory (i.e., the causality-contextualized word embedding) and non-parametric memory (i.e., a classic data retrieval-based engine) for language generation (Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D., 2020). However, the shortage of maintenance text data may hinder the exploitation of this approach. Therefore, a NLP augmentation strategy could be helpful (Bayer, M., Kaufhold, M.-A., Buchhold, B., Keller, M., Dallmeyer, J., and Reuter, C., 2021), although the larger the data analyzed, the greater the chance that spurious correlations dominate the results and lead to erroneous conclusions (Dima, A., Lukens, S., Hodkiewicz, M., Sexton, T., and Brundage, M. P., 2021). Alternatively, fine-tuning a bigger pre-trained language model, which has become the de facto standard for doing transfer learning in NLP, could also be advantageous (Li, J., Tang, T., Zhao, W. X., and Wen, J.-R., 2021). Finally, the deployment of the presented approach to a different railway PHM asset such as the Passenger Door System may reveal further CI insights into the integration of FMMEA with ROX (Dinmohammadi, F., Alkali, B., Shafiee, M., Bérenguer, C., and Labib, A., 2016), and with the increased availability of diverse SCM, a Graph Neural Network could expect to learn a truly holistic troubleshooting system at the train level (Bronstein, M. M., Bruna, J., Cohen, T., Velickovic, P., 2021).

ACKNOWLEDGMENT

We would like to show our gratitude to Prof. Francesc Alías, Dr. Jonathan Brown, and Dr. Eduardo Di-Santi 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

- Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, J., Schölkopf, B., Wüthrich, M., and Bauer, S. (2020). CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning. *arXiv:2010.04296 [cs.RO]*, 1–18.
- Alaa, A. M., Weisz, M., and van der Schaar, M. (2017). Deep Counterfactual Networks with Propensity-Dropout. *Proc. of the 34th International Conference on Machine Learning*, 1–6.
- Almeida, F., and Xexéo, G. (2019). Word Embeddings: A Survey. *arXiv:1901.09069 [cs.CL]*, 1–10.
- Ansaldi, S. M., Agnello, P., Pirone, A., and Vallerotonda, M. R. (2021). Near Miss Archive: A Challenge to Share Knowledge among Inspectors and Improve Seveso Inspections. *Sustainability*, 13(8456), 1–21.
- Ansari, F. (2020). Cost-based text understanding to improve maintenance knowledge intelligence in manufacturing enterprises. *Computers and Industrial Engineering*, 141(106319).
- Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant Risk Minimization. *arXiv:1907.02893 [stat.ML]*, 1–31.
- Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N. (2017). Prognostics and Health Management for Maintenance Practitioners - Review, Implementation and Tools Evaluation. *International Journal of Prognostics and Health Management*, 8(60), 1–31.
- Bahadori, M. T., and Heckerman, D. E. (2021). Debiasing Concept Bottleneck Models with a Causal Analysis Technique. *Proc. of the International Conference on Learning Representations*, 1–11.
- Bahadori, M. T., Chalupka, K., Choi, E., Chen, R., Stewart, W. F., and Sun, J. (2017). Causal Regularization. *arXiv:1702.02604 [cs.LG]*, 1–18.
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *Proc. of the International Conference on Learning Representations*, 1–15.
- Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. *arXiv:1801.09536 [cs.CL]*, 1–26.
- Barocas, S., Selbst, A. D., and Raghavan, M. (2019). The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. *arXiv:1912.04930 [cs.CY]*,

- 1–17.
- Bayer, M., Kaufhold, M.-A., Buchhold, B., Keller, M., Dallmeyer, J., and Reuter, C. (2021). Data Augmentation in Natural Language Processing: A Novel Text Generation Approach for Long and Short Text Classifiers. *arXiv:2103.14453 [cs.CL]*, 1–20.
- Bengio, Y. (2017). The Consciousness Prior. *arXiv:1709.08568 [cs.LG]*, 1–7.
- Bronstein, M. M., Bruna, J., Cohen, T., Velickovic, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478 [cs.LG]*, 1–156.
- Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, 27, 42–46.
- Brundage, M. P., Weiss, B. A., and Pellegrino, J. (2020). Summary Report: Standards Requirements Gathering Workshop for Natural Language Analysis. *National Institute of Standards and Technology Advanced Manufacturing Series*, 100(30), 1–50.
- Cai, H., Zheng, V. H., and Chang, K. C.-C. (2018). A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 30, 1616–1637.
- Camacho-Collados, J., and Pilehvar, M. T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research*, 63, 743–788.
- Chen, W., Grangier, D., and Auli, M. (2015). Strategies for Training Large Vocabulary Neural Language Models. *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, 1975–1985.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 1–14.
- Conrad, S. (2019). Register in English for Academic Purposes and English for Specific Purposes. *Register Studies*, 1(1), 168–198.
- Crawshaw, M. (2020). Multi-Task Learning with Deep Neural Networks: A Survey. *arXiv:2009.09796 [cs.LG]*, 1–43.
- Dima, A., Lukens, S., Hodkiewicz, M., Sexton, T., and Brundage, M. P. (2021). Adapting natural language processing for technical text. *Applied AI Letters*, 2(e33), 1–11.
- Dinmohammadi, F., Alkali, B., Shafiee, M., Bérenguer, C., and Labib, A. (2016). Risk Evaluation of Railway Rolling Stock Failures Using FMECA Technique: A Case Study of Passenger Door System. *Urban Rail Transit*, 2(3–4), 128–145.
- DuBay, W. H. (2004). The Principles of Readability. *Impact Information*, 1–77.
- Ebrahimipour, V., Rezaie, K., and Shokravi, S. (2010). An ontology approach to support FMEA studies. *Expert Systems with Applications*, 37(1), 671–677.
- Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. *arXiv:2009.05451 [cs.CL]*, 1–12.
- 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.
- Gelman, A., and Imbens, G. (2013). *Why ask why? Forward causal inference and reverse causal questions* (Tech. Rep. No. 19614). National Bureau of Economic Research.
- Gelman, A., and Vehtari, A. (2020). What are the most important statistical ideas of the past 50 years? *arXiv:2012.00174 [stat.ME]*, 1–19.
- Goth, G. (2016). Deep or Shallow, NLP Is Breaking Out. *Communications of the ACM*, 59(3), 13–16.
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2222–2232.
- Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). A Survey of Learning Causality with Data: Problems and Methods. *ACM Computing Surveys*, 53(4).
- Gutmann, M., and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Proc. of the 13th International Conference on Artificial Intelligence and Statistics*, 297–304.
- Hancock, J. T., and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 7(28), 1–41.
- Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. *Proc. of the 34th International Conference on Machine Learning*, 1–10.
- Hastings, E. M., Sexton, T., Brundage, M. P., and Hodkiewicz, M. (2019). Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 1–7.
- Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *arXiv:1907.07271 [stat.ME]*, 1–76.
- ISO. (2003). *Condition monitoring and diagnostics of machines – General guidelines on data interpretation and diagnostics techniques* (Tech. Rep. No. 13379:2003(E)). International Organization for Standardization.
- Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., and

- Daumé III, H. (2014). A Neural Network for Factoid Question Answering over Paragraphs. *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 633–644.
- Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs.CL]*, 1–11.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs.LG]*, 1–30.
- Karray, M. H., Ameri, F., Hodkiewicz, M., and Louge, T. (2019). ROMAIN: Towards a BFO compliant reference ontology for industrial maintenance. *Applied Ontology*, 14(2), 155–177.
- Le, Q., and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proc. of the 31st International Conference on Machine Learning*, 1–9.
- Leao, B. P., Fitzgibbon, K. T., Puttini, L. C., and de Melo, G. P. B. (2008). Cost-Benefit Analysis Methodology for PHM Applied to Legacy Commercial Aircraft. *Proc. of IEEE Aerospace Conference*, 1–13.
- LeCun, Y. and Bengio, Y., and Hinton, G. E. (2015). Deep Learning. *Nature*, 521, 436–444.
- Lee, K., Firat, O., Agarwal, A., Fannjiang, C., and Sussillo, D. (2018). Hallucinations in Neural Machine Translation. *Proc. of the 32th Conference on Neural Information Processing Systems*, 1–18.
- Leuenberger, H., Puchkov, M., and Schneider, B. (2013). Right, First Time Concept and Workflow. A Paradigm Shift for a Smart & Lean Six-sigma Development. *Swiss Pharma*, 35(3), 3–16.
- Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2, 302–308.
- Levy, O., and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Proc. of the 27th International Conference on Neural Information Processing Systems*, 2, 2177–2185.
- Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3, 211–225.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proc. of the International Conference on Neural Information Processing Systems*, 1–18.
- Li, J., Tang, T., Zhao, W. X., and Wen, J.-R. (2021). Pre-trained Language Models for Text Generation: A Survey. *arXiv:2105.10311 [cs.CL]*, 1–9.
- Li Y., and Yang T. (2018). Word Embedding for Understanding Natural Language: A Survey. *Guide to Big Data Applications. Studies in Big Data*, 26, 83–104.
- Li, Y.-H., Wang, Y.-D., and Zhao, W.-Z. (2009). Bogie Failure Mode Analysis for Railway Freight Car Based on FMECA. *Proc. of the 8th International Conference on Reliability, Maintainability and Safety*, 5–8.
- Liu, Q., Kusner, M. J., and Blunsom, P. (2020). A Survey on Contextual Embeddings. *arXiv:2003.07278 [cs.CL]*, 1–13.
- Liu, Z., Jia, Z., Vong, C.-M., Han, W., Yan, C., and Pecht, M. (2018). A Patent Analysis of Prognostics and Health Management (PHM) Innovations for Electrical Systems. *IEEE Access*, 6, 18088–18107.
- Maguire, P., Mulhall, O., Maguire, R., and Taylor, J. (2015). Compressionism: A Theory of Mind Based on Data Compression. *Proc. of the 11th International Conference on Cognitive Science*, 294–299.
- Mathew, S., Das, D., Rossenberger, R., and Pecht, M. (2008). Failure Mechanisms Based Prognostics. *Proc. of the International Conference on Prognostics and Health Management*, 1–6.
- Mihalcea, R., and Radev, D. (2011). *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proc. of Workshop at the International Conference on Learning Representations*, 1–12.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Proc. of the Conference on Neural Information Processing Systems*, 1–9.
- Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proc. of the North American Chapter of the Association for Computational Linguistics*, 746–751.
- Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C. (2021). Representation Learning via Invariant Causal Mechanisms. *Proc. of the International Conference on Learning Representations*, 1–12.
- Mnih, A., and Teh, Y. W. (2012). A Fast and Simple Algorithm for Training Neural Probabilistic Language Models. *Proc. of the 29th International Conference on Machine Learning*, 1–8.
- Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *Proc. of the Conference on Fairness, Accountability, and Transparency*, 1–13.
- Nastase, V., Mihalcea, R., and Radev, D. (2015). A survey of graphs in natural language processing. *Natural Language Engineering*, 21(5), 665–697.
- Naveed, A., Li, J., Saha, B., Saxena, A., and Vachtsevanos, G. (2012). A Reasoning Architecture for Expert Troubleshooting of Complex Processes. *Proc. of the Annual Conference of the Prognostics and Health Management*

- Society*, 1–8.
- Navinchandran, M., Sharp, M. E., Brundage, M. P., and Sexton, T. B. (2019). Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 1–11.
- Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96–146.
- Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54–60.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365 [cs.CL]*, 1–15.
- Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019). Language Models as Knowledge Bases? *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 1–11.
- Polenghi, A., Roda, I., Macchi, M., and Pozzetti, A. (2022). Ontology-augmented Prognostics and Health Management for shopfloor-synchronised joint maintenance and production management decisions. *Journal of Industrial Information Integration*, 27(100286).
- PwC. (2019). *It's time for a consumer-centred metric: introducing 'return on experience'*. *Global Consumer Insights Survey* (Tech. Rep. No. 512587-2019). PricewaterhouseCoopers International Limited.
- Rehman, Z., and Kifor, C. V. (2016). An Ontology to Support Semantic Management of FMEA Knowledge. *International Journal of Computers Communications & Control*, 11(4), 507–521.
- Salakhutdinov, R., and Hinton, G. (2009). Semantic hashing. *International Journal of Approximate Reasoning*, 50, 969–978.
- 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.
- Sexton, T. B., and Brundage, M. P. (2019). Nestor: A Tool for Natural Language Annotation of Short Texts. *Journal of Research of National Institute of Standards and Technology*, 124(124029), 1–5.
- Sexton, T., Hodkiewicz, M., Brundage, M. P., and Smoker, T. (2018). Benchmarking for Keyword Extraction Methodologies in Maintenance Work Orders. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 1–10.
- Shalit, U., Johansson, F. D., and Sontag, D. (2016). Estimating individual treatment effect: generalization bounds and algorithms. *arXiv:1606.03976 [stat.ML]*, 1–20.
- Sharma, A., and Kiciman, E. (2020). DoWhy: An End-to-End Library for Causal Inference. *arXiv:2011.04216 [stat.ME]*, 1–5.
- Sharp, M. E., Sexton, T. B., and Brundage, M. P. (2017). Semi-Autonomous Labeling of Unstructured Maintenance Log Data for Diagnostic Root Cause Analysis. *Proc. of the International Conference Advances in Production Management Systems*, 1–8.
- Shen, Z., Cui, P., Kuang, K., Li, B., and Chen, P. (2018). Causally Regularized Learning with Agnostic Data Selection Bias. *Proc. of ACM Multimedia Conference*, 1–9.
- Smetkowska, M., and Mrugalska, B. (2018). Using Six Sigma DMAIC to Improve the Quality of the Production Process: A Case Study. *Procedia – Social and Behavioral Sciences*, 238, 590–596.
- Stampe, D. W. (2008). Towards A Causal Theory of Linguistic Representation. *Midwest Studies in Philosophy*, 2(1), 42–63.
- Su, Y., Awadallah, A. H., Wang, M., and White, R. H. (2018). Natural Language Interfaces with Fine-Grained User Interaction: A Case Study on Web APIs. *Proc. of the 41th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1–10.
- Suter, R., Miladinović, D., Schölkopf, B., and Bauer, S., (2019). Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness. *Proc. of the 36th International Conference on Machine Learning*, 1–10.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv:1409.3215 [cs.CL]*, 1–9.
- Tan, L., Zhang, H., Clarke, C. L. A., and Smucker, M. D. (2015). Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings. *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, 657–661.
- Tan, S., Zhou, Z., Xu, Z., and Li, P. (2019). On Efficient Retrieval of Top Similarity Vectors. *Proc. of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5236–5246.
- Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A. (2020). Natural Language Processing Advancements By Deep Learning: A Survey. *arXiv:2003.01200 [cs.CL]*, 1–21.
- Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X. (2021). Integrated Multiple-Defect Detection and Evaluation of Rail Wheel Tread Images using Convolutional Neural Networks. *International Journal of Prognostics and Health Management*, 12(1), 1–19.
- Vlontzos, A., Kainz, B., and Gilligan-Lee, C. M. (2021).

- Estimating the probabilities of causation via deep monotonic twin networks. *arXiv:2109.01904 [cs.LG]*, 1–10.
- Wang, C., and Sennrich, R. (2020). On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 3544–3552.
- Wang, Z., and Wang, H. (2016). Understanding Short Texts. *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, 1–4.
- Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H. (2019). Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings. *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 5740–5753.
- 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.
- Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2020). A Survey on Causal Inference. *arXiv:2002.02770 [stat.ME]*, 1–38.
- Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., and Jiang, M. (2020). A Survey of Knowledge-Enhanced Text Generation. *arXiv:2010.04389 [cs.CL]*, 1–44.
- Zanette, D. H., and Montemurro, M. A. (2005). Dynamics of Text Generation with Realistic Zipf's Distribution. *Journal of Quantitative Linguistics*, 12(1), 29–40.
- Zheng, M., Marsh, J. K., Nickerson, J. V., and Kleinberg, S. (2020). How causal information affects decisions. *Cognitive Research: Principles and Implications*, 5(6), 1–24.

BIOGRAPHIES

Alexandre Trilla graduated from La Salle University of Barcelona with a M.Sc. in Electronics and Telecommunications

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.

5.5 Journal Article 3 (2023)

Unsupervised Probabilistic Anomaly Detection over Nominal Subsystem Events on a Hierarchical Variational Autoencoder (2023)

This contribution develops a method to discover and diagnose anomalies in massive operational data for subsystem event signals using a Variational Autoencoder built with convolutional layers and extended hierarchically with a Multilayer Perceptron. Additionally, its results yield interesting opportunities for designing Intrusion Detection Systems in the context of cybersecurity.

This article was first published on May 2023 in the International Journal of Prognostics and Health Management (Trilla, A., Mijatovic, N., and Vilasis-Cardona, X., 2023), and then presented on September 2023 at the 2nd Causal AI Conference in New York City.

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

Alexandre Trilla^{1,3}, Nenad Mijatovic², and Xavier Vilasis-Cardona³

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

Alexandre Trilla et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

<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

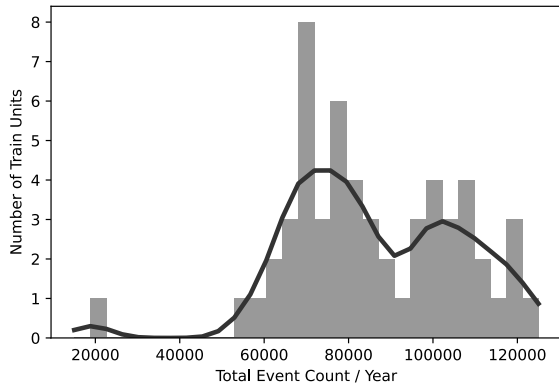


Figure 1. Histogram of the total event count per train unit, showing two main modes as humps in the kernel density estimation.

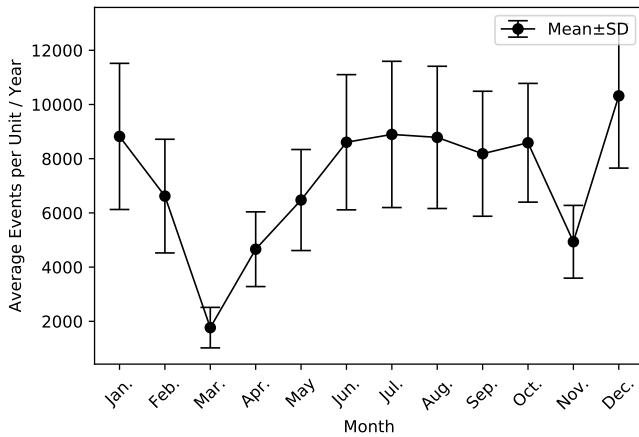


Figure 2. Monthly evolution of event counts.

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

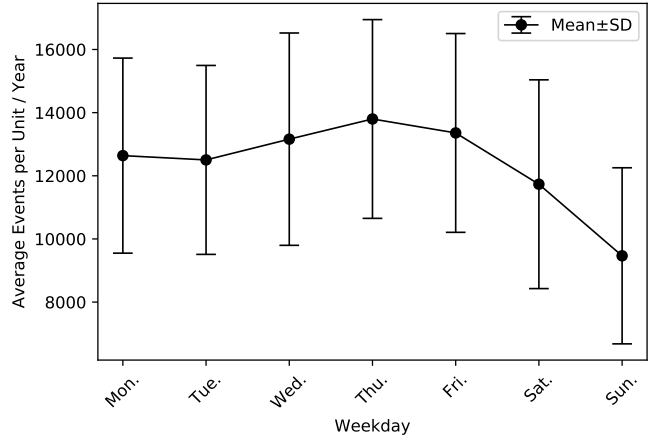


Figure 3. Weekly evolution of event counts.

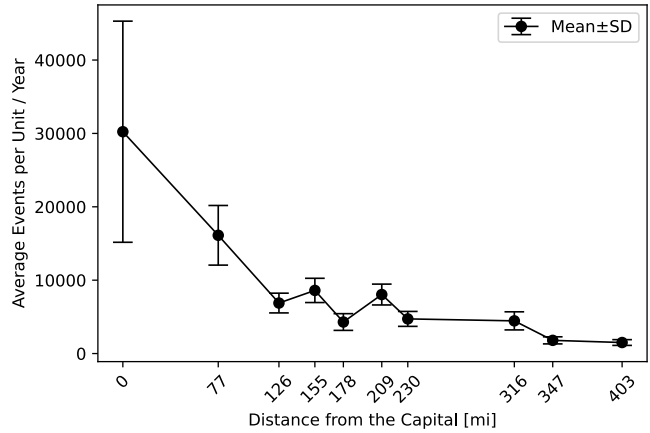


Figure 4. Evolution of event counts given the location.

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-

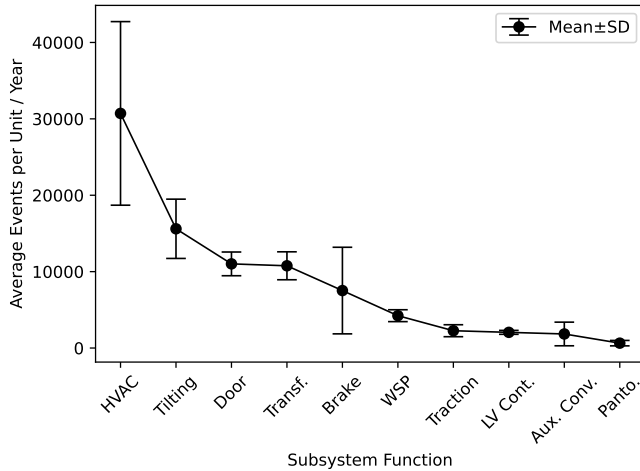


Figure 5. Ranked total event counts given the subsystems.

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,

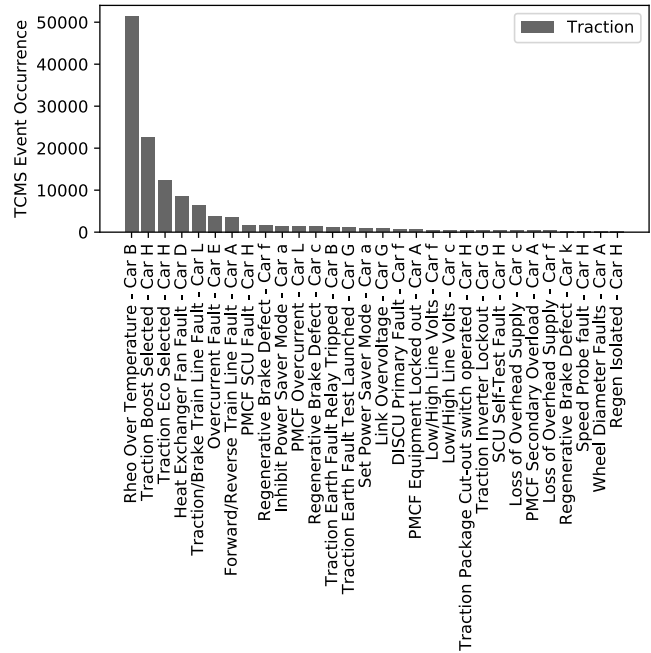


Figure 6. Histogram of the top 30 frequency-ranked events for the Traction subsystem.

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

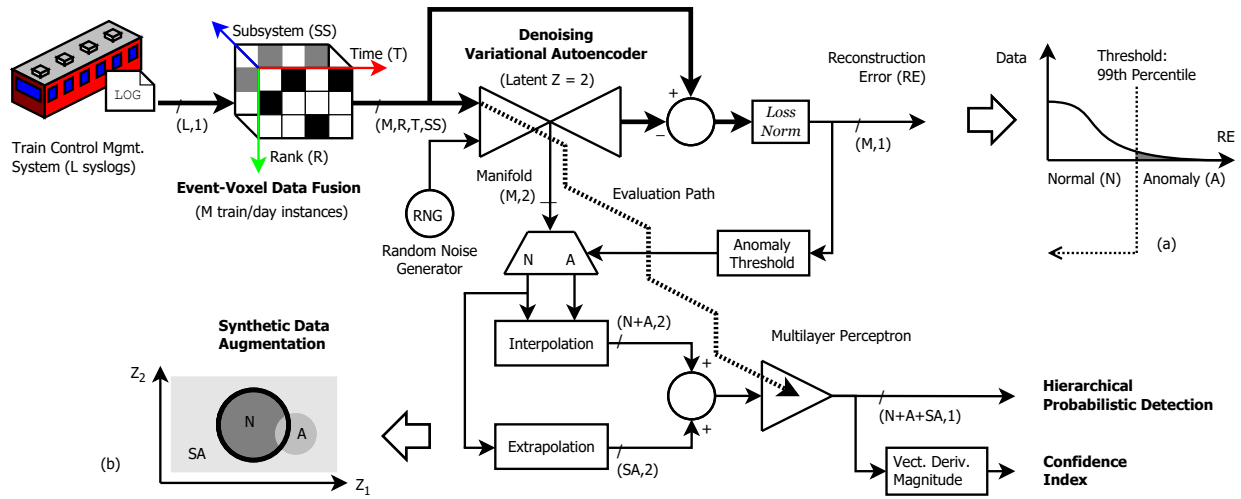


Figure 7. Diagram of the proposed Anomaly Detection framework. Plot (a) depicts the expected distribution of the Reconstruction Error. Plot (b) depicts the expected representation on the augmented latent space. This design is mostly focused on training the solution. Regarding its industrial deployment, 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.

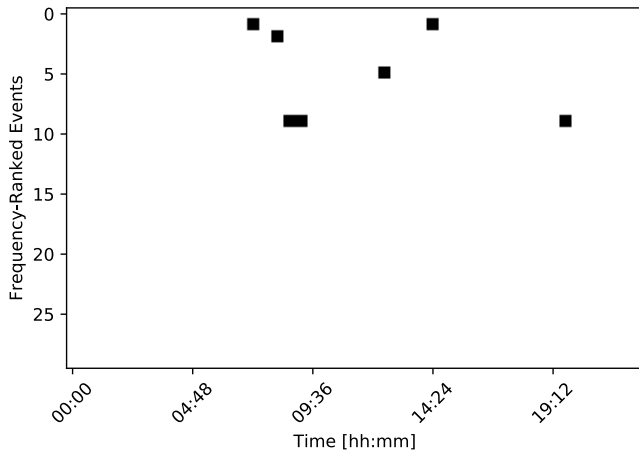


Figure 8. Chart representation of a Normal condition pattern.

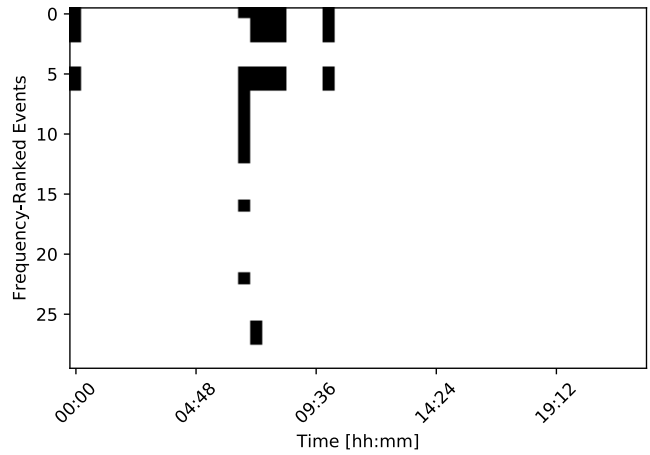


Figure 9. Chart representation of an Anomaly condition pattern.

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 in-

tegrated 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

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

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 Percep-

tron (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).

$$p_A(z) = g(W_H(g(W_I z))) \quad (2)$$

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.

$$CI(z) = 1.0 - \min(\|\nabla p_A(z)\|, 1.0) \quad (3)$$

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).

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad (4)$$

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.)

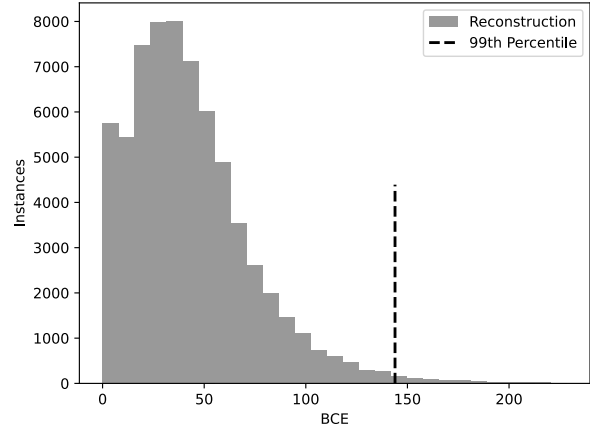


Figure 10. Histogram of the Binary Cross Entropy (BCE) Reconstruction Error along with the 99th percentile threshold. The plot roughly matches the expected distribution of this Loss, see Figure 7(a).

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

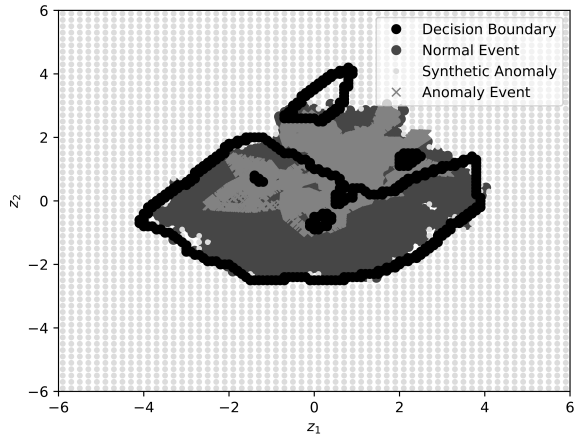


Figure 11. Tessellation of the latent manifold on the learned bidimensional embedding $z = (z_1, z_2)$. The probabilistic anomaly decision boundary is shown at $p_A(z) = 0.5$, which is the random guess on a dichotomic classification problem. Note that while the latent space is continuous, the evaluation 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).

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

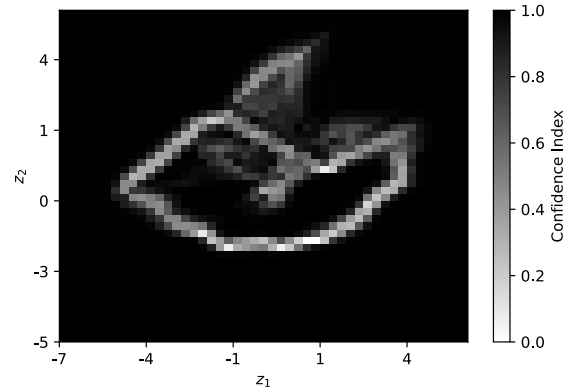


Figure 12. Confidence Index shown on the latent manifold related to Figure 11.

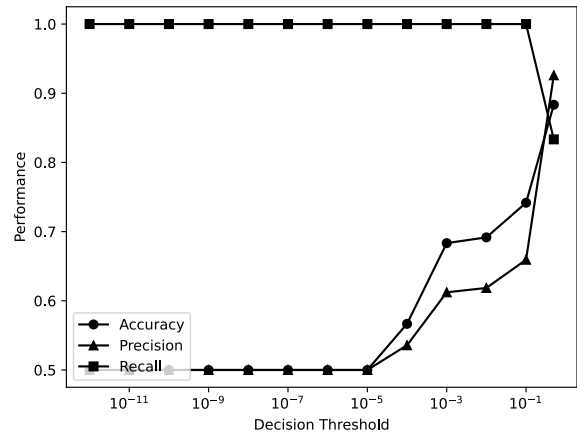


Figure 13. Precision and Recall curves driven by the sensitivity of the Decision Threshold. Accuracy is also shown here only for reference as the total rate of correct classifications.

hindsight, though, simplifications to the proposed approach could now be found, but these seem unlikely to 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 (Rodríguez 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	$p_A(\mathbf{N})[\mu/\sigma]$	$p_A(\mathbf{A})[\mu/\sigma]$	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 from 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

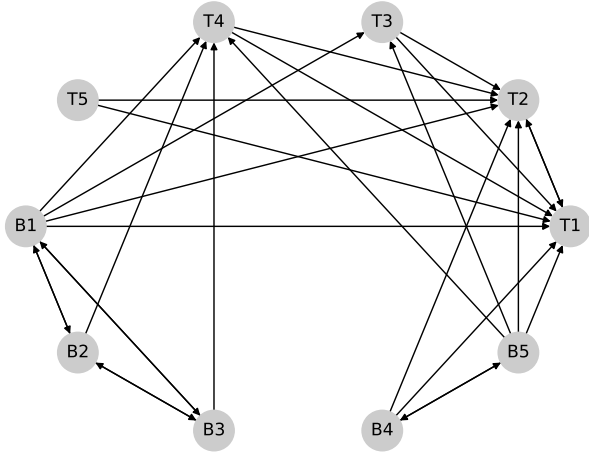


Figure 14. Causal graph for the Locomotion environment, i.e., including the Traction (T) and Brake (B) subsystems. Node name code: {Subsystem}{Rank}. See Table 3 for further details. Arrows indicate event association from cause to effect.

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

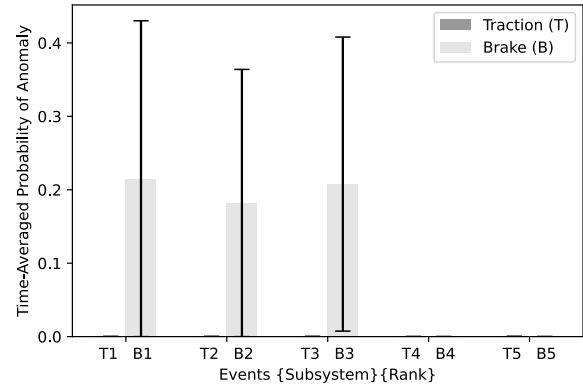


Figure 15. Sensitivity analysis on the Locomotion environment. See Table 3 for further details. Assuming Normality for the day-level average distributions, bar heights indicate their mean values, and whiskers indicate one standard deviation. All the visually imperceptible bars actually have a negligible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

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 Stiefelwagen, 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: Lessons 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 Stiefelwagen, 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 Fonladosa, 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., Dhimi, 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, oper-

ations, 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.

DISCUSSION AND CONCLUSIONS

The generation of consciousness through a nonlinear neural net that tries to solve the binding problem to provide more effective computations strikes me as unconvincing and almost insulting.
– James A. Anderson (1996)

THE implementation of predictive maintenance is part of a complex business and corporate transformative process. In the foreseeable future, predictive maintenance will be done in conjunction with some more traditional maintenance approaches (UITP, 2020).

The research work described in this dissertation has focused on proving that the Deep Learning technology exhibits the features that make it suitable for implementing railway predictive maintenance solutions. Through several studies and investigations, comprising different data characteristics (e.g., nominal and parametric variables) and objectives (e.g., diagnosis and prognosis), the main conclusion is that Deep Learning is an approach that is consistent with other studies, and which adds value to the predictive maintenance products driven by its superior performance and flexibility. While the specific developed solutions cannot be exactly replicated because the data used was private, the complexity of the approaches and the details of their descriptions were deemed to be sufficiently accurate to guarantee their reproducibility in a similar setting. These characteristics would support the

advice for other researchers in PHM (and possibly other fields and industries) to adopt Deep Learning for their applied industrial research.

Deep Learning (DL) has also pushed the frontiers of knowledge, and new limitations have ensued. The following sections describe some of the current weaknesses of Deep Learning, as well as some avenues of future improvement. This chapter is organized as follows: Section 6.1 deals with the topic of Interpretability and Explainable AI. Section 6.2 introduces the process of decision making. Section 6.3 outlines some aspects of the industrialization of DL-based solutions. Section 6.4 introduces the science of causality as an line of research to improve the limitations of DL. Section 6.5 discuss how the published work addresses the identified challenges and opportunities, and Section 6.6 concludes the dissertation.

6.1 Interpretability and Explainability

BLACK-BOX MODEL

Predictive maintenance applications are increasingly complex, with interactions between many components. Black-box models such as the recent ones based on Deep Learning are popular approaches due to their unprecedented performance in predictive accuracy. However, the lack of model explainability or interpretability may manifest itself in a lack of trust to address PHM problems.

EXPLAINABILITY
INTERPRETABILITY

The challenges of interpretable Machine Learning for PHM include: the fact that explanation methods interpreting black-box models may show black-box behavior themselves, the non-consistent use of terminology, and the inclusion of domain knowledge (Vollert, S., Atzmueller, M., and Theissler, A., 2021). One recent successful approach in PHM to shed light into these hidden behaviors is the application of online rule learning algorithms to explain when the black-box models predict rare events (Ribeiro, R. P., Mastelini, S. M., Davari, N., Aminian, E., Veloso, B., and Gama, J., 2023). Obviously, other more general approaches such as LIME (based on local linear approximations) or SHAP (based on coalitional game theory) may also be explored (Molnar, C., 2019).

Different approaches have been investigated in this research to shed light on the internal behavior of the developed solutions. The most common technique has been the Sensitivity Analysis, which has been applied on the input data (Trilla, A., Fernández, V., and Cabré, X., 2020; Trilla, A., Mijatovic, N., and Vilasis-Cardona, X., 2023), on the model expressiveness (Trilla, A., Miralles, D., and Fernández, V., 2020), and on the intermediate probabilistic representations (Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X., 2021). Additionally, the learned preprocess-

ing has been interpreted through the templates of the matched filters, and its overall performance has also been tested against data corruption to better understand its limitations (Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X., 2021). Finally, the extraction of the structure that binds the variables through their relationships has also been used to explain the root causes of specific events (Trilla, A., Mijatovic, N., and Vilasis-Cardona, X., 2023).

6.2 Decision Making

Many important problems involve making decisions under uncertainty, including PHM. When designing automated decision-making systems or decision-support systems, it is important to account for the various sources of uncertainty when making or recommending decisions (Kochenderfer, M. J., Wheeler, T. A., and Wray, K. H., 2022). Solutions for managing such uncertainty may be based on a deep map between measurements and optimal operation performance scores (Rodriguez Garcia, G., Michau, G., Einstein, H. H., and Fink, O., 2021), and the exploitation of the predicted Remaining Useful Life for optimizing business processes (Wesendrup, K., and Hellingrath, B., 2020). Overall, consensus over multiple independent solutions must be sought: learning to combine different predictions through an ensemble is a means to reduce the uncertainty when making decisions (Gupta, N., Smith, J., Adlam, B., and Mariet, Z., 2022).

Such ensembles, which may also augment the input data with their specific context, have been explored extensively in this research as a means to increase the robustness of the decisions that may be derived from the results obtained with the developed solutions (Trilla, A., Fernández, V., and Cabré, X., 2020; Trilla, A., Mijatovic, N., and Vilasis-Cardona, X., 2022, 2023; Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X., 2021). Finally, the field of probability has been regarded as a key element to deal with the uncertainty that is inevitably linked to typical unstructured data such as images (Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X., 2021) and text (Trilla, A., Mijatovic, N., and Vilasis-Cardona, X., 2022), but also with the sparseness of random events in time (Trilla, A., Mijatovic, N., and Vilasis-Cardona, X., 2023).

DECISION-MAKING

UNCERTAINTY

CONSENSUS

ENSEMBLE

CONTEXT

6.3 Industrialization

This section focuses on two main issues involved in the broad expansion of Deep Learning solutions: their technical debt and their energy consump-

TECHNICAL DEBT tion. On the one hand, technical debt deals with the long term costs incurred by moving quickly in software engineering (Sculley, D., Holt, G., *et al.*, 2015). These include entanglement, hidden feedback loops, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns. To counter this adverse effect, testing and monitoring are two key considerations for ensuring the production-readiness of Deep Learning systems, and also for reducing their technical debt (Breck, E., Cai, S., Nielsen, E., Salib, M., and Sculley, D., 2017).

CARBON FOOTPRINT On the other hand, the computations required for Deep Learning research have been doubling every few months, resulting in an estimated 300,000x increase from 2012 to 2018, and showing a surprisingly large carbon footprint (Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O., 2020). The massive computation required to obtain the impressive results that Deep Learning has yielded is costly both financially, due to the price of specialized hardware and electricity or cloud compute time, and to the environment, as a result of the non-renewable energy used to fuel modern tensor processing hardware (Strubell, E., Ganesh, A., and McCallum, A., 2020). As a result, efficiency is a novel criterion that is being increasingly considered in the evaluation of these solutions.

This doctoral research has specifically tackled the impact of energy consumption through reduced data transmission using neural data compression (Trilla, A., Miralles, D., and Fernández, V., 2020). This approach has been deployed on an Industrial Internet of Things solution that captures vibration degradation patterns. Finally, reporting the details of a cloud implementation in Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X. (2021) was positively regarded by the journal reviewers that helped improve the publication. This is considered as a step forward in the reduction of technical debt.

6.4 Causal Inference

ASSOCIATIONAL INTERVENTIONAL COUNTERFACTUAL CAUSAL MODEL Deep Learning has succeeded primarily by showing that certain questions or tasks that were thought to be difficult were in fact not (Pearl, J., and Mackenzie, D., 2019). Deep Learning and other modern data mining tools are now placed on the bottom rung of the Ladder of Causation, i.e., the associational layer (Goldberg, L., 2019). More involved reasoning strategies, such as the interventional and counterfactual approaches, can provide finer insights into the data. However, they also need explicit assumptions on the processes that generated the data. In this sense, Deep Learning and Causal Models have recently found a sweet spot to enrich one another (Zečević, M.,

Dhami, D. S., Veličković, P., and Kersting, K., 2021; Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E., 2021). In fact, the concept of “Causal AI” has just appeared in the Gartner’s hype cycle as one of the most promising new technologies (Gartner, 2022).

In PHM, one of the specific diagnosis areas where Causal Inference finds a good match is Root Cause Analysis (RCA). In general, the diagnostic capability of operational tech systems is supported by having subject matter experts identify causes by the patterns in historical data sets (Apps, J., 2019). This is again a case of observational data managed through associations, and thus confusion persists by wrongly linking causality with correlation for RCA (Salehi, R., and Duan, S., 2023). In railway maintenance, though, an easy way to identify the root cause of an abnormal behavior is to swap some parts of a system with a neighboring system and see if the problem swaps too (Turgis, F., Audier, P., Nemoz, V., and Marion, R., 2022). This is a clear example of how actions, or interventions, are used to discover causal implications.

ROOT CAUSE

RCA requires reliable, explainable and understandable models such as Bayesian Networks (BN) for performing tasks like condition prediction (Pourret, O., 2008). BN can also be learned with incomplete data and in a supervised or unsupervised way, which is very useful because the collection of labeled data is costly and sometimes impossible in PHM (Monvoisin, M., Leray, P., and Ritou, M., 2021). Furthermore, BN can also model the uncertainty in the parameters with a distribution, which can be useful to complement the assumptions made in the failure analyses (Mascherona, R., Bellani, L., Compare, M., Trucco, R., Zio, E., 2020).

BAYESIAN NETWORK

UNCERTAINTY

The increasingly growing area of causality within AI circles has been specifically addressed in this research. In Trilla, A., Mijatovic, N., and Vilasis-Cardona, X. (2022), a distributed representation of linguistic features and causality was developed for RCA purposes. A troubleshooting solution was proposed which treated the text-based records as Bags of Words, and modeled their causal entailment in the diagnosis direction while conditioning on the common project context to adjust for any confounding factor. Finally, causality was also adopted in Trilla, A., Mijatovic, N., and Vilasis-Cardona, X. (2023), where the discovery of structure in a multivariate environment was introduced to enhance the interpretability and explainability of the data.

6.5 Updated Challenges and Opportunities

Section 3.1.3 summarized the main challenges and opportunities identified in the Deep Learning literature, which can be taken for the current research gaps in the technical arena. This section revisits them from the perspective of the published works shown in Chapter 5, and tries to see how they helped to move the PHM field forward. Table 6.1 charts the challenges and opportunities with the publications to provide a quick overview of the scope of the covered research.

Regarding the original challenges and opportunities, Table 6.1 aggregates Model Selection and Benchmarking, and Data Scarcity and Augmentation, Industrial Data Characteristics and Data Analysis.

The contribution CP1, which developed a multivariate regression refinement of pantograph carbon strip degradation in time considering the impact of the season, addressed the different seasonal environments as distinct domains, and evaluated their sensitivity. This solution was industrialized in a monitoring product based on computer vision technologies.

The contribution CP2, which studied and benchmarked the compression performance of mechanical vibration signals using different regularization strategies, was industrialized in an edge computing IoT product.

The contribution JA1, which developed a minimum viable product for wheel tread diagnosis, used image augmentations to train several classification and regression models, and put a lot of emphasis on its robustness and explainability to increase the confidence in its assessment.

The contribution JA2, which approached the acquisition of causal information from text, leveraged different project environments, developed a fundamental word model for troubleshooting, and addressed its interpretability through the resulting learned distributed representation.

The contribution JA3, which jointly tackled the detection and diagnosis objectives using event signals, managed diverse system settings, exploited several augmentation strategies, and focused on its interpretability through different representation spaces, including a real-valued manifold and a discrete graph.

The big gap in the chart is the lacking contribution on the Real-Time Realization challenge. In all the works conducted in this research dissertation, keeping the solution in continuous improvement entails retraining the models. Thus, performance drifts have to be monitored externally, and the relearning needs to be conducted periodically.

Challenge/Opportunity	CP1 (2020)	CP2 (2020)	JA1 (2021)	JA2 (2022)	JA3 (2023)
Cross-domain Prediction	Seasonality			Context (business, system, issue)	Intersubsystem Diversity
Industrial Data Scarcity and Augmentation			Affine transformations and additive noise		Translation (time-shift) and Synthetic Minority Oversampling Technique
Model Selection and Benchmarking	Regression ensemble	Regularized Autoencoders	Classification and regression streams	Separate word embedding and language model	Reconstruction-based detection and graph representation
Interpretability and Explainability	Sensitivity		Filter analysis, manifold plot, sensitivity, input corruption robustness	Ontology, manifold plot	Manifold plot, causal graph, sensitivity
Real-Time Realization					

Table 6.1 Aggregated challenges and opportunities along with the Conference Papers (CP) and Journal Articles (JA) that addressed them, also showing the publication years in brackets.

6.5.1 Railway Fleet Planning

The vast majority of this dissertation has been focused on the tasks of detecting anomalies and diagnosing the health condition of the assets. However, for such advances to be transformed into value-added actions, the maintenance planning of the fleet needs to be observed. The challenge for PHM is that predictive maintenance may require very short-term schedule changes which may also affect the operation of the trains.

SCHEDULING

In a railway network, predictive maintenance scheduling for trains aims to maximize the system reliability and availability such that sufficient capacity for the passenger demand in each route of the network is satisfied. In a liberalized railway market, where the business is split between the infrastructure, the rolling stock, and the operator (Glover, J., 2013), one of the most interesting approaches to tackle this challenge is to model it as a centralized game theory problem (Rokhforoz, P., and Fink, O., 2021). In this setting, the central system seeks to maximize its reward, which it gets from the price per passenger for each route, and to minimize the operation and deterioration cost of the trains, which are composed of several wagons that keep their degradation as private health information. To solve this problem, the central system needs to design a mechanism that induces a non-cooperative game among the wagon agents, the solution of which is conceptualized by a Nash Equilibrium (Nash, J., 1951), which in turn results in the most effective maintenance schedule.

PLANNING

According to this formulation of the problem, the opportunity of PHM to improve the planning of the fleet is introduced through the prognosis, specifically addressed as the estimation of the distribution of the Remaining Useful Life. In this sense, contribution CP1 provides a point estimate of the future degradation along with the uncertainty of the error (assuming Normality), and contribution JA1 provides a confidence indicator and a classification label, which is translated into granting a grace period (up to one week) for further testing on the shop floor.

GAME THEORY

Finally, it is to note that the conceptualization of railway planning as a game theory problem introduced in Rokhforoz, P., and Fink, O. (2021) has interesting connections with the field of Causal Inference, especially regarding the canonical econometric model relating price and demand through structural equations (Pearl, J., 2000).

6.5.2 Technical Language Processing in Retrospect

The 2022 NLP contribution JA2 has been the most controversial article given the feedback received, which probably signals the novelty of that re-

search. The intersection of predictive maintenance, language, and causality, has not been easily received by the PHM community as a matter of fact. In hindsight, the arguably unconvincing results are likely to be the result of the fresh training of the models using the scarce data that characterizes a predictive maintenance setting. Learning from that experience, now the problem could alternatively be addressed through fine-tuning already-trained global word embeddings and large language models, but in that case the intricate essence of causality would not be learned directly from the entailment of the text, which may raise some concerns.

To the best of our knowledge, the introduction of language processing in PHM was first popularized when the National Institute of Standards and Technology created the Technical Language Processing Community of Interest¹ in 2020. This group helped researchers in PHM and language processing gravitate toward a common interest in this area. Concurrently, the traditional community of Natural Language Processing (NLP) and computational linguistics held the first workshop on Causal Inference and NLP (Feder, A., Keith, K. A., *et al.*, 2021), highlighting the challenge of extracting linguistic features from text that also represent causal effects.

The contribution JA2 approached the application of NLP on PHM tackling the causal challenge of root cause analysis troubleshooting through DL, making use of a learned distributed linguistic representation of causality. To this day, no other work is known to precede and to follow this line of research. However, an interesting alternative focus is to appear two years after JA2. Valcamonico, D., Baraldi, Zio, E., Decarli, L., Crivellari, A., and La Rosa, L. (2024) developed a more traditional (i.e., non-DL-based) approach with a Bayesian Network that had an expert-driven deconfounded causal structure and leveraged keyword spotting. Given that this approach smartly avoids the issue of common causes that induce confounding in the structure, the conditional probabilities computed with the model can be attributed a causal meaning.

6.6 On a Final Note...

The applied research work conducted in this dissertation has tried to build a convincing case for proving that the Deep Learning technology is suitable for implementing railway predictive maintenance solutions. After the rigorous evaluation of the proposed methods compared to leading alternatives over different data sets and problem scenarios, it can be concluded that Deep

¹<https://www.nist.gov/el/technical-language-processing-community-interest>

Learning adds value to several predictive maintenance products, driven by its superior performance and flexibility.

Looking ahead, two of the areas where Deep Learning has been particularly successful, i.e., Computer Vision and Natural Language Processing, are now being enhanced with causal reasoning technology to improve visual interpretations (Liu, Y., Wei, Y.-S., Yan, H., Li, G.-B., and Lin, L., 2022; Zhang, K., Sun, Q., Zhao, C., and Tang, Y., 2023) and language models (Kıcıman, E., Ness, R., Sharma, A., and Tan, C., 2023; Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez, F., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B., 2023). Therefore, causality is a clear promising field to be adopted to continue the pursuit of knowledge acquisition in AI.

BIBLIOGRAPHY

- Aimar, M., and Somà, A. (2018). Study and results of an onboard brake monitoring system for freight wagons. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 232(5):1277–1294.
- Alemi, A., Corman, F., Pang, Y., and Lodewijks, G. (2019). Reconstruction of an informative railway wheel defect signal from wheel–rail contact signals measured by multiple wayside sensors. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 233(1):49–62.
- Altman, B., and Odetunde, S. (2022). Variable rate sanding improves braking. *International Railway Journal*, 62(2):35–37.
- Antunes, P., Pombo, J., Ambrósio, J., Rebelo, J., Santos, J. (2022). Supporting Railway Electrification Projects with an Integrated Pantograph-Catenary Dynamic Analysis Tool. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Apps, J. (2019). Augmented Decision-Making: When Data Replaces Experience. *Uptime*, pages 32–33.
- Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N. (2017). Prognostics and Health Management for Maintenance Practitioners - Review, Implementation and Tools Evaluation. *International Journal of Prognostics and Health Management Special Issue on Railway Systems & Mass Transportation*, 8(3):1–31.

- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *Proc. of the International Conference on Learning Representations*, pages 1–15.
- Barr, A., and Smith, K. (2022). Batteries and digitalisation key in Hitachi’s growth strategy. *International Railway Journal*, 62(10):42.
- Barraza, J. F., Droguett, E. L., and Martins, M. R. (2021). Embedded Feature Importance Determination Technique for Deep Neural Networks Based Prognostics and Health Management. *Proc. of the 31st European Safety and Reliability Conference*, pages 1494–1501.
- Barrow, K. (2018a). Bridging the big data gap. *International Railway Journal*, 58(7):40–44.
- Barrow, K. (2018b). Can AI take customer service to the next level? *International Railway Journal*, 58(8):66.
- Barrow, K. (2018c). Guarding rail against evolving threats. *International Railway Journal*, 58(4):20–25.
- Barrow, K. (2019a). Mobility as a Service, the end of the road for urban car ownership? *International Railway Journal*, 59(8):24–27.
- Barrow, K. (2019b). Smart Cities, Making the connection. *International Railway Journal*, 59(2):36–40.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. of the National Academy of Sciences*, 116(32):15849–15854.
- Ben Taleb Ali, M., Schrevre, T., Pedron, A., Blanvillain, G., and Auditeau, G. (2022). Intelligent shocks detector on catenary infrastructure. *Proc. of the World Congress on Railway Research*, pages 1–7.
- Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–71.
- Berner, J., Grohs, P., Kutyniok, G., and Petersen, P. (2022). The Modern Mathematics of Deep Learning. *Mathematical Aspects of Deep Learning*, pages 1–111.
- Bešinović, N., De Donato, L., Flammini, F., *et al.* (2022). Artificial Intelligence in Railway Transport: Taxonomy, Regulations, and Applications. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14011–14024.

- Blome, C., Kargoll, B., and Wernz, J. (2022). Is consistent real-time data the secret to improving customer satisfaction? *International Railway Journal*, 62(9):48–49.
- Boyes, H., Hallaq, B., Cunningham, J., and Watson, T. (2018). The industrial internet of things (IIoT): An analysis framework. *Computers in Industry*, 101:1–12.
- Breck, E., Cai, S., Nielsen, E., Salib, M., and Sculley, D. (2017). The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction. *Proc. of the IEEE International Conference on Big Data*, pages 1–10.
- Briginshaw, D. (2017). Automation spurs operational rethink. *International Railway Journal*, 57(10):31–34.
- Briginshaw, D. (2019). FRMCS: next-generation train radio begins to take shape. *International Railway Journal*, 59(7):32–37.
- Briginshaw, D. (2020a). Strong leadership needed to achieve CO2 reduction targets. *International Railway Journal*, 60(2):4.
- Briginshaw, D. (2020b). Will ERTMS ever reach critical mass in Europe? *International Railway Journal*, 60(2):20–24.
- Briginshaw, D. (2021). Leaders highlight post Covid-19 challenges. *International Railway Journal*, 61(2):34.
- Briginshaw, D. (2022a). Clear vision needed for rail to grow. *International Railway Journal*, 62(1):10–12.
- Briginshaw, D. (2022b). Europe’s ERTMS dream enters a new era. *International Railway Journal*, 62(8):38–41.
- Britz, D. (2020). Deep Learning’s Most Important Ideas - A Brief Historical Review. *Journal of Petroleum Technology*, pages 1–16.
- Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, 27:42–46.
- Brundage, M. P., Weiss, B. A., and Pellegrino, J. (2020). Summary Report: Standards Requirements Gathering Workshop for Natural Language Analysis. *National Institute of Standards and Technology Advanced Manufacturing Series*, 100(30):1–50.

- Bruni, S., Mistry, P. J., Johnson, M. S., *et al.* (2022). A vision for a lightweight railway wheelset of the future. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 236(10):1179–1197.
- Brunner, S., Frischknecht-Gruber, C. M.-L., Reif, M., and Senn, C. W. (2022). Deep Gaussian Mixture Model - A Novelty Detection Method for Time Series. *Proc. of the 32nd European Safety and Reliability Conference*, pages 1291–1298.
- Burroughs, D. (2018). The quiet revolution in noise abatement. *International Railway Journal*, 58(11):44–45.
- Burroughs, D. (2019a). Big data culture, an evolving strategy for IoT in rail. *International Railway Journal*, 59(7):38–41.
- Burroughs, D. (2019b). Combating the cybersecurity threat. *International Railway Journal*, 59(10):40–43.
- Burroughs, D. (2019c). The future of intelligence is artificial. *International Railway Journal*, 59(9):46–51.
- Burroughs, D. (2019d). Training responds to changing workforce. *International Railway Journal*, 59(12):34–35.
- Burroughs, D. (2020). Preparing to face the digital threat. *International Railway Journal*, 60(9):40–44.
- Burroughs, D. (2021a). Managing network-wide RCF. *International Railway Journal*, 61(12):33–36.
- Burroughs, D. (2021b). The latest in rail industry innovation. *International Railway Journal*, 61(9):41–45.
- Burroughs, D. (2022a). Battery traction recharges decarbonisation fight. *International Railway Journal*, 62(10):32–35.
- Burroughs, D. (2022b). Is technology key to restoring passenger trust? *International Railway Journal*, 62(9):42–46.
- Burroughs, D. (2022c). Making rail construction sustainable from conception to completion. *International Railway Journal*, 62(7):40–45.
- Burroughs, D. (2022d). Pandemic fallout affects volumes and revenues. *International Railway Journal*, 62(1):14–15.

- Burroughs, D. (2023). Rail's digital vulnerabilities worry cyber experts. *International Railway Journal*, 63(2):32–35.
- Burroughs, D., Wust, D., and Wust, J. (2023). AI opens door to intelligent rolling stock. *International Railway Journal*, 63(4):30–33.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *Proc. of the 37th International Conference on Machine Learning*, 149:1597–1607.
- Cheng, H., Cao, Y., Wang, J., Zhang, W., and Zeng, H. (2020). A preventive, opportunistic maintenance strategy for the catenary system of high-speed railways based on reliability. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 234(10):1149–1155.
- Clark, M. (2022). What patent trends tell us about rail innovation. *International Railway Journal*, 62(7):38–39.
- Clinnick, R. (2021a). DB and Siemens demonstrate automated S-Bahn. *International Railway Journal*, 61(12):30–32.
- Clinnick, R. (2021b). Hydrogen: how realistic is it for rail traction? *International Railway Journal*, 61(10):24–26.
- Clinnick, R. (2021c). LRV automation progresses in Potsdam. *International Railway Journal*, 61(10):37–39.
- Clinnick, R. (2022a). Rail helps Ukraine evacuation and aid efforts. *International Railway Journal*, 62(4):5–6.
- Clinnick, R. (2022b). The new Karlsruhe Model: data sharing. *International Railway Journal*, 62(5):37–38.
- Compare, M., Baraldi, P., and Zio, E. (2019). Challenges to IoT-enabled predictive maintenance for industry 4.0. *IEEE Internet of Things Journal*, 7(5):4585–4597.
- Cooney, N. (2020). British group develops autonomous wheel control system. *International Railway Journal*, 60(2):41–43.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314.

- da Rosa, T. G., de Andrade Melani, A. E., Kashiwagi, F. N., de Carvalho Michalski, M. A., de Souza, G. F. M., de Oliveira Salles, G. M., and Rigoni, E. (2022). Data Driven Fault Detection in Hydroelectric Power Plants based on Deep Neural Networks. *Proc. of the 32nd European Safety and Reliability Conference*, pages 1235–1242.
- De Simone, L., Caputo, E., Cinque, M., Galli, A., Moscato, V., Russo, S., Cesaro, G., Criscuolo, V., and Giannini, G. (2023). LSTM-based failure prediction for railway rolling stock equipment. *Expert Systems with Applications*, 222(119767).
- Derosa, S., Frøseth, G. T., Lau, A., and Rönquist, A. (2022). Vehicle-infrastructure interaction monitoring from train in traffic. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Derosa, S., Nåvik, P., Collina, A., Bucca, G., and Rönquist, A. (2020). A heuristic wear model for the contact strip and contact wire in pantograph - Catenary interaction for railway operations under 15 kV 16.67 Hz AC systems. *Wear*, 456–457(203401):1–8.
- Derosa, S., Nåvik, P., Collina, A., Bucca, G., and Rönquist, A. (2021). Contact point lateral speed effects on contact strip wear in pantograph-catenary interaction for railway operations under 15 kV 16.67 Hz AC systems. *Wear*, 486–487(204103):1–9.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs.CL]*, pages 1–16.
- DeVol, N., Saldana, C., and Fu, K. (2022). Evaluating Image Classification Deep Convolutional Neural Network Architectures for Remaining Useful Life Estimation of Turbofan Engines. *International Journal of Prognostics and Health Management*, 13:1–12.
- Ding, X., Chen, H., Zhang, X., Han, J., and Ding, G. (2022). RepMLPNet: Hierarchical Vision MLP with Re-parameterized Locality. *Proc. of the IEEE / CVF Computer Vision and Pattern Recognition Conference*, pages 1–13.
- Dong, K., Romanov, I., McLellan, C., and Esen, A. F. (2022). Recent text-based research and applications in railways: A critical review and future trends. *Engineering Applications of Artificial Intelligence*, 116(105435):1–19.

- Duda, R. O., Hart, P. E., and Stork, D. G., editor (2001). *Pattern Classification*. Wiley-Interscience.
- Eisenberger, D., and Fink, O. (2017). Assessment of maintenance strategies for railway vehicles using Petri-nets. *Transportation Research Procedia*, 27:205–214.
- Elsken, T., Metzen, J. H., and Hutter, F. (2019). Neural Architecture Search: A Survey. *Journal of Machine Learning Research*, 20:1–21.
- Englbrecht, M. (2022). Improving braking offers capacity increase. *International Railway Journal*, 62(2):38–39.
- Entezami, M., Roberts, C., Weston, P., Stewart, E., Amini, A., and Pappalias, M. (2020). Perspectives on railway axle bearing condition monitoring. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 234(1):17–31.
- Fararoy, S., and Allan, J. (1995). On-line condition monitoring of railway equipment using neural networks. *Proc. of the IEE Colloquium on Advanced Condition Monitoring Systems for Railways*, 5162514.
- Farzad, A., and Gulliver, A. (2020). Unsupervised log message anomaly detection. *ICT Express*, 6:229–237.
- Feder, A., Keith, K. A., *et al.* (2021). Proceedings of the First Workshop on Causal Inference and NLP. *Transactions of the Association for Computational Linguistics*.
- Feder, A., Keith, K. A., *et al.* (2022). Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.
- 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.
- Fink, O., Zio, E., and Weidmann, U. (2015a). A Classification Framework for Predicting Components' Remaining Useful Life Based on Discrete-Event Diagnostic Data. *IEEE Transactions on Reliability*, 64(3):1049–1056.

- Fink, O., Zio, E., and Weidmann, U. (2015b). Development and Application of Deep Belief Networks for Predicting Railway Operation Disruptions. *International Journal of Performability Engineering*, 11(2):121–134.
- Fink, O., Zio, E., and Weidmann, U. (2015c). Fuzzy Classification With Restricted Boltzman Machines and Echo-State Networks for Predicting Potential Railway Door System Failures. *IEEE Transactions on Reliability*, 64(3):861–868.
- Frankle, J., and Carbin, M. (2019). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *Proc. of the International Conference on Learning Representations*, pages 1–42.
- Garrido Martínez-Llop, P., Sanz Bobi, J. de D., and Huera Plaza, A. (2022). Application of neural networks for the prediction of railway bearing failures. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 236(10):1147–1153.
- Gartner (2022). Hype Cycle for Emerging Tech.
- Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proc. of the 13th International Conference on Artificial Intelligence and Statistics, Proc. of Machine Learning Research*, 9:249–256.
- Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proc. of the 14th International Conference on Artificial Intelligence and Statistics, Proc. of Machine Learning Research*, 15:315–323.
- Glover, J. (2013). *The Principles of Railway Operation*. Ian Allan Publishing.
- Goldberg, L. (2019). The Book of Why - A review by Lisa R. Goldberg. *Notices of the American Mathematical Society*, 66(7):1093–1098.
- Goldthorpe, P., and Desmet, A. (2018). Denoising autoencoder anomaly detection for correlated data. *Proc. of the Fourth European Conference of the Prognostics and Health Management Society*, pages 1–6.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. *arXiv:1406.2661 [stat.ML]*, pages 1–9.
- Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6):1789–1819.

- Groom, S., Doshi-Keeble, F., and Williams, P. (2022). GOTCHA Wheelset Damage Management on the UK Class 390 Pendolino. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Gupta, N., Smith, J., Adlam, B., and Mariet, Z. (2022). Ensembles of Classifiers: a Bias-Variance Perspective. *Transactions on Machine Learning Research*, pages 1–23.
- Hameed, R. (2021). Batteries could solve ac-dc transfer issue. *International Railway Journal*, 61(10):30–33.
- Hamming, R. W. (1986). *You and Your Research*. Stripe Press.
- Hapgood, F. (1993). *Up the Infinite Corridor: MIT and the Technical Imagination*. Addison-Wesley Publishing Company.
- He, D. (2023). *Security and Communication Networks - Special Issue on Cryptographic Schemes and Protocols for Artificial Intelligence*. Hindawi.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Heidarysafa, M., Kowsari, K., Barnes, L., and Brown, D. (2018). Analysis of Railway Accidents’ Narratives Using Deep Learning. *Proc. of the 17th IEEE International Conference on Machine Learning and Applications*.
- Helland, P. (2020). The Best Place to Build a Subway. Building projects despite (and because of) existing complex systems. *ACM Queue*, pages 1–9.
- Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580 [cs.NE]*, pages 1–18.
- Hornik, K. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2:359–366.
- Hosseinpour, F., Ahmed, I., Baraldi, P., Behzad, M., Zio, E., and Lewitschnig, H. (2022). An Unsupervised Method for Anomaly Detection in Multi-Stage Production Systems Based on LSTM Autoencoders. *Proc.*

- of the 32nd European Safety and Reliability Conference*, pages 1346–1352.
- Hu, Y., Miao, X., Si, Y., Pan, E., and Zio, E. (2022). Prognostics and health management: A review from the perspectives of design, development and decision. *Reliability Engineering and System Safety*, 217(108063):1–15.
- Ioffe, S., and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs.LG]*, pages 1–11.
- Jahan, S. and Mahmud, A. S. (2015). What Is Capitalism? *Finance & Development*, 52(2):44–45.
- Jarillo, J. M., Moreno, J., Alfi, S., *et al.* (2021). Novel technology concepts and architecture for on-board condition-based monitoring of railway running gear: The RUN2Rail vision. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 235(5):616–630.
- Jernelv, I. L., Hjelme, D. R., Matsuura, Y., and Aksnes, A. (2020). Convolutional neural networks for classification and regression analysis of one-dimensional spectral data. *arXiv:2005.07530 [cs.LG]*, pages 1–18.
- Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez, F., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B. (2023). CLadder: Assessing Causal Reasoning in Language Models. *Proc. of the Advances in Neural Information Processing Systems 36*, pages 1–27.
- 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*, pages 1–14.
- Kawaguchi, T., Sueki, T., Kitagawa, T., Nishimura, M., and Abe, H. (2019). Wheel/rail noise above 10 kHz generated on a gently curved track. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Khan, S., and Yairi, T. (2018). A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 107:241–265.
- Kıçıman, E., Ness, R., Sharma, A., and Tan, C. (2023). Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *arXiv:2305.00050 [cs.AI]*, pages 1–42.

- Kilian, K., Kilian, M., Mazur, V., and Phelan, J. (2016). Rethinking reliability engineering using machine vision systems. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 230(3):1006–1014.
- Kingma, D. P., and Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. *Proc. of the 3rd International Conference for Learning Representations*, pages 1–15.
- Kobayashi, S., Koyama, T., and Harada, S. (2022). Hybrid Simulation of Pantograph/Catenary Systems using High-Speed Pantograph Testing Machine. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Kochenderfer, M. J., Wheeler, T. A., and Wray, K. H. (2022). *Algorithms for Decision Making*. The MIT Press.
- Komiyama, J., and Maehara, T. (2018). A Simple Way to Deal with Cherry-picking. *arXiv:1810.04996 [stat.ME]*, pages 1–25.
- Kour, R., Aljumaili, M., Karim, R., and Tretten, P. (2019). eMaintenance in railways: Issues and challenges in cybersecurity. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 233(10):1012–1022.
- Kour, R., Karim, R., and Thaduri, A. (2020). Cybersecurity for railways - A maturity model. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 234(10):1129–1148.
- Kramer, M. A. (1992). Autoassociative Neural Networks. *Computers and Chemical Engineering*, 16(4):313–328.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25:1097–1105.
- Kumari, J., Karim, R., Thaduri, A., and Castano, M. (2022). Augmented asset management in railways - Issues and challenges in rolling stock. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 236(7):850–862.
- Kutyniok, G. (2022). The Mathematics of Artificial Intelligence. *arXiv:2203.08890 [cs.LG]*, pages 1–16.

- Le Glatin, N., and Clarke, P. (2021). A feasibility study towards the conceptual development of a real-time digital twin to reduce dwell time variations on the Thameslink route (COF-DSP-06). Technical Report COF-DSP-06, Rail Safety and Standards Board.
- LeCun, Y. and Bengio, Y., and Hinton, G. E. (2015). Deep Learning. *Nature*, 521:436–444.
- Lee, W.-J. (2017). Anomaly Detection and Severity Prediction of Air Leakage in Train Braking Pipes. *International Journal of Prognostics and Health Management Special Issue on Railway Systems & Mass Transportation*, 8(3):1–12.
- Leuenberger, H., Puchkov, M., and Schneider, B. (2013). Right, First Time Concept and Workflow. A Paradigm Shift for a Smart & Lean Six-sigma Development. *Swiss Pharma*, 35(3):3–16.
- Liu, B., Ghazel, M., and Toguyéni, A. (2018). On-the-Fly and Incremental Technique for Fault Diagnosis of Discrete Event Systems Modeled by Labeled Petri Nets. *Asian Journal of Control*, 20(1):1–13.
- Liu, H., Dai, Z., So, D. R., and Le, Q. V. (2021). Pay Attention to MLPs. *arXiv:2105.08050 [cs.LG]*, pages 1–15.
- Liu, J., Schmid, F., Li, K., and Zheng, W. (2021). A knowledge graph-based approach for exploring railway operational accidents. *Reliability Engineering & System Safety*, 207(107352).
- Liu, Y., Wei, Y.-S., Yan, H., Li, G.-B., and Lin, L. (2022). Causal Reasoning Meets Visual Representation Learning: A Prospective Study. *Machine Intelligence Research*, 19(6):485–511.
- Liu, Y. Z., Zou, Y. S., Wu, Y., Zhang, H. Y., and Ding, G. F. (2022). A novel abnormal detection method for bearing temperature based on spatiotemporal fusion. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 236(3):317–333.
- Lortie, M., and Holmes, E. (2014). Automated Inspection Technologies: a new paradigm for preventive maintenance programs. *Proc. of the Rail Conference*, pages 1–13.
- Man, T. (2018). Condition monitoring improves asset and network performance. *International Railway Journal*, 58(11):40–43.

- Mascherona, R., Bellani, L., Compare, M., Trucco, R., Zio, E. (2020). Enhancements of Reliability Centered Maintenance analysis and its application to the railway industry. *Proc. of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*, pages 1–7.
- Matsumoto, T., Nishidouzono, K., Fukaya, F., Koga, S., Nakamura, H., and Kameda, M. (2022). Implementing Contact Line Monitoring System. *Proc. of the World Congress on Railway Research*, pages 1–6.
- McKinsey (2017). The rail sector’s changing maintenance game. *Digital*, pages 1–24.
- Melas-Kyriazi, L. (2021). Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet. *arXiv:2105.02723 [cs.CV]*, pages 1–4.
- Miciek, R. (2019). Failure to Launch: Why IoT Projects Fail to Get Started. *Uptime*, pages 26–27.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proc. of Workshop at the International Conference on Learning Representations*, pages 1–12.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Proc. of the Conference on Neural Information Processing Systems*, pages 1–9.
- Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proc. of the North American Chapter of the Association for Computational Linguistics*, pages 746–751.
- Miller, K., and Dubrawski, A. (2019). System-Level Predictive Maintenance: Review of Research Literature and Gap Analysis. *arXiv:2005.05239 [cs.AI]*, pages 1–24.
- Mittermayr, P., Schmid, R., Zottl, W., Betterle, E., Occioni, G. (2019). OBAL measurement system - Wheel profile measurement at train speeds up to 250 km/h. *Proc. of the 12th World Congress on Railway Research*, pages 1–6.

- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. *Proc. of the Conference on Neural Information Processing Systems - Deep Learning Workshop*, pages 1–9.
- Molnar, C. (2019). *Interpretable Machine Learning*. Leanpub.
- Monvoisin, M., Leray, P., and Ritou, M. (2021). Unsupervised co-training of Bayesian networks for the diagnosis of machining spindle. *Proc. of the 31st European Safety and Reliability Conference*, pages 1990–1997.
- Moyne, J., Balta, E. C., Kovalenko, I., Faris, J., Barton, K., and Tilbury, D. M. (2020). A Requirements Driven Digital Twin Framework: Specification and Opportunities. *IEEE Access*, 8:107781–107801.
- Nakamura, R. (2019). Clearing up the airwaves. *International Railway Journal*, 59(12):40–41.
- Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019). Deep Double Descent: Where Bigger Models and More Data Hurt. *arXiv:1912.02292 [cs.LG]*, pages 1–24.
- Nash, J. (1951). Non-cooperative games. *Annals of mathematics*, pages 286–295.
- Oberhuber, H., Neuhold, J., Orta Roca, J., Brandl, D., and Schönhuber, B. (2021). Maintaining metro track through milling. *International Railway Journal*, 61(12):37–39.
- O’Hanlon, T. (2019). Redefining Asset Performance Management. *Uptime*, pages 10–14.
- Oomen, M. A., Bosman, R., and Lugt, P. M. (2017). Characterization of Friction and Wear Behavior of Friction Modifiers used in Wheel-Rail Contacts. *International Journal of Prognostics and Health Management Special Issue on Railway Systems & Mass Transportation*, 8(3):1–13.
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Pearl, J., and Mackenzie, D. (2019). *The Book Of Why: The New Science of Cause and Effect*. Penguin Books Ltd.

- Pennacchi, P., Bruni, S., Chatterton, S., Borghesani, P., Ricci, R., Marinis, D., Didonato, A., and Unger-Weber, F. (2011). A Test Rig for the Condition-Based Maintenance Application of the Traction Chain of Very High Speed Trains. *Proc. of the World Congress on Railway Research*, pages 1–12.
- Pincioli, L., Baraldi, P., and Zio, E. (2023). Maintenance optimization in Industry 4.0. *Reliability Engineering & System Safety*, 109204.
- Poupart-Lafarge, H., and Smith, K. (2021). Alstom aims to win global innovation battle with Bombardier Transportation takeover. *International Railway Journal*, 61(2):14–17.
- Pourret, O. (2008). Introduction to Bayesian networks. *Bayesian Networks: A Practical Guide to Applications*, pages 1–13.
- Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *Proc. of the International Conference on Learning Representations*, pages 1–16.
- Raghu, M., and Schmidt, E. (2020). A Survey of Deep Learning for Scientific Discovery. *arXiv:2003.11755 [cs.LG]*, pages 1–48.
- Rebelo, J., Pombo, J., Antunes, P., Santos, J., Magalhães, H., and Ambrósio, J. (2022). Advanced Studies for Improved Current Collection Performance at Catenary Gradients. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Reeve, J. (2019). Demanding Excellence from Your Asset Management System. *Uptime*, pages 36–40.
- Remadna, I., Terrissa, L. S., Ayad, S., and Zerhouni, N. (2021). RUL Estimation Enhancement using Hybrid Deep Learning Methods. *International Journal of Prognostics and Health Management*, 12:1–19.
- Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., and Wang, X. (2020). A Comprehensive Survey of Neural Architecture Search: Challenges and Solutions. *ACM Computing Surveys*, 37(4):111:1–111:33.
- Rezaeianjouybari, B., and Shang, Y. (2020). Deep learning for prognostics and health management: State of the art, challenges, and opportunities. *Measurement*, 163(107929).

- Ribeiro, R. P., Mastelini, S. M., Davari, N., Aminian, E., Veloso, B., and Gama, J. (2023). Online Anomaly Explanation: A Case Study on Predictive Maintenance. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Communications in Computer and Information Science*, 1753:383–399.
- Rodenbeck, A., and Clinnick, R. (2022). Digitalisation key to unlocking capacity. *International Railway Journal*, 62(7):50.
- Rodriguez Garcia, G., Michau, G., Einstein, H. H., and Fink, O. (2021). Decision support system for an intelligent operator of utility tunnel boring machines. *Automation in Construction*, 131(103880):1–12.
- Rokhforoz, P., and Fink, O. (2021). Hierarchical multi-agent predictive maintenance scheduling for trains using price-based approach. *Computers & Industrial Engineering*, 159(107475).
- Romanchikov, A., and Smith, K. (2022). A Train Control revolution. *International Railway Journal*, 62(2):26–29.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.
- Salehi, R., and Duan, S. (2023). Signal Abstraction for Root Cause Identification of Control Systems Malfunctions in Connected Vehicles. *International Journal of Prognostics and Health Management*, 14:1–8.
- 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*, pages 918–923.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.
- Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12):54–63.
- Scordamaglia, D. (2019). Digitalisation in railway transport. Technical Report PE 635.528, European Parliament.
- Sculley, D., Holt, G., *et al.* (2015). Hidden Technical Debt in Machine Learning Systems. *Proc. of the Advances in Neural Information Processing Systems 28*, pages 1–9.

- Seisenberger, M., ter Beek, M. H., Fan, X., Ferrari, A., Haxthausen, A. E., James, P., Lawrence, A., Luttk, B., van de Pol, J., and Wimmer, S. (2022). Safe and Secure Future AI-Driven Railway Technologies: Challenges for Formal Methods in Railway. *Leveraging Applications of Formal Methods, Verification and Validation. Practice. ISoLA 2022. Lecture Notes in Computer Science*, 13704:246–268.
- Selvanathan, B., Nistala, S. H., Runkana, V., Desai, S. J., and Agarwal, S. (2023). Ensemble Deep Learning for Detecting Onset of Abnormal Operation in Industrial Multi-component Systems. *International Journal of Prognostics and Health Management*, 14:1–14.
- Shabtay, L., Fournier-Viger, P., Yaari, R., and Dattner, I. (2021). A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data. *Information Sciences*, 553:353–375.
- Shackleton, P., Sztrauch, K., Eickhoff, B., and Bevan, A. (2022). Using wheel impact load detector data for the identification of vehicle defects in freight wagons. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Shi, L., Zhu, Y., Zhang, Y., and Su, Z. (2021). Fault Diagnosis of Signal Equipment on the Lanzhou-Xinjiang High-Speed Railway Using Machine Learning for Natural Language Processing. *Complexity*, 2021(9126745):1–13.
- Shift2Rail (2020). Multi-Annual Action Plan. Technical Report HI-01-20-118-EN-N, Office of the European Union.
- Si, J., Shi, H., and Yang, J. (2022). Evaluation of Chinese freight train bearing condition based on spatiotemporal feature extraction. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 236(9):1047–1057.
- Smith, K. (2018). Centralised diagnostics aids transition to predictive maintenance. *International Railway Journal*, 58(6):46–48.
- Smith, K. (2021). Policymakers must see rail as a solution to the climate crisis. *International Railway Journal*, 61(11):4.
- Smith, K. (2022a). Energy crisis presents tough choices for Europe’s railways. *International Railway Journal*, 62(10):4.

- Smith, K. (2022b). Rail's skills challenges need urgent attention. *International Railway Journal*, 62(12):4.
- Smith, K. (2022c). Railway researchers urged to make an impact. *International Railway Journal*, 62(7):4.
- Smith, K. (2023). Intense pressure on SMEs threatens rail project delivery. *International Railway Journal*, 63(1):4.
- Song, B., Zhang, Z., Qin, Y., Liu, Y., and Hu, H. (2022). Quantitative analysis of freight train derailment severity with structured and unstructured data. *Reliability Engineering & System Safety*, 224(108563).
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.
- Stefanakis, G., Nyktari, V., Papaioannou, A., and Askitopoulou, H. (2020). Hippocratic concepts of acute and urgent respiratory diseases still relevant to contemporary medical thinking and practice: a scoping review. *BMC Pulmonary Medicine*, 20(165):1–7.
- Stenström, C., Al-Jumaili, M., and Parida, A. (2015). Natural language processing of maintenance records data. *International Journal of CO-MADEM*, pages 1–5.
- Stone, V. M. (2008). The auto-associative neural network - a network architecture worth considering. *Proc. of the 2008 World Automation Congress*, pages 1–4.
- Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and Policy Considerations for Modern Deep Learning Research. *Proc. of the 34th AAAI Conference on Artificial Intelligence*, pages 13693–13696.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv:1409.3215 [cs.CL]*, pages 1–9.
- Sutton, R. S. (2019). The Bitter Lesson. Technical report, University of Alberta.
- Syeda, K. N., Shirazi, S. N., Naqvi, S. A. A., Parkinson, H. J., and Bamford, G. (2019). Big Data and Natural Language Processing for Analysing Railway Safety. *Big Data and Natural Language Processing for Analysing Railway Safety: Analysis of Railway Incident Reports*, pages 781–809.

- Tamssaouet, F., Nguyen, K. T. P., Medjaher, K., and Orhard, M. (2021). Combination of Long Short-Term Memory and Particle Filtering for Future Uncertainty Characterization in Failure Prognostic. *Proc. of the 31st European Safety and Reliability Conference*, pages 275–281.
- Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A Survey on Deep Transfer Learning. *Proc. of the 27th International Conference on Artificial Neural Networks*, pages 1–10.
- Tang, R., De Donato, L., Bešinović, N., Flammini, F., Goverde, R. M. P., Lin, Z., Liu, R., Tang, T., Vittorini, V., and Wang, Z. (2022). A literature review of Artificial Intelligence applications in railway systems. *Transportation Research Part C: Emerging Technologies*, 140(103679):1–25.
- Teodoro, I. P., Ribeiro, D. F., Botari, T., Martins, T. S., and Santos, A. A. (2019). Fast simulation of railway pneumatic brake systems. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 233(4):420–430.
- Thompson, I. (2022). Installing a Condition Monitoring System onto an Existing Metro Fleet. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., *et al.* (2021). MLP-Mixer: An all-MLP Architecture for Vision. *arXiv:2105.01601 [cs.CV]*, pages 1–16.
- Touvron, H., Bojanowski, P., Caron, M., *et al.* (2021). ResMLP: Feed-forward networks for image classification with data-efficient training. *arXiv:2105.03404 [cs.CV]*, pages 1–17.
- Trausmuth, A., Schmid, R., Dinhl, G., and Badisch, E. (2022). Experimental simulation of wheel-rail contact for optimized lifetime of the infrastructure in the rail network. *Proc. of the World Congress on Railway Research*, pages 1–6.
- Trilla, A., and Cabré, X. (2018). Determining the Equivalent Conicity for Railway Wheelset Maintenance with Deep Ensembles. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 10(1):1–6.
- Trilla, A., and Gratacòs, P. (2013). Condition based maintenance on board. *Chemical Engineering Transactions Journal*, (33):733–738.

- Trilla, A., and Gratacòs, P. (2016). Maintenance of bogie components through vibration inspection with intelligent wireless sensors: a case-study of axle-boxes and wheel-sets using the Empirical Mode Decomposition technique. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, (230):1408–1414.
- Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X. (2021). Integrated Multiple-Defect Detection and Evaluation of Rail Wheel Tread Images using Convolutional Neural Networks. *International Journal of Prognostics and Health Management*, 12(1):1–19.
- Trilla, A., Dersin, P., and Cabré, X. (2018). Estimating the Uncertainty of Brake Pad Prognostics for High-speed Rail with a Neural Network Feature Ensemble. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 10(1):1–7.
- Trilla, A., Fernández, V., and Cabré, X. (2020). Enhancing Railway Pantograph Carbon Strip Prognostics with Data Blending through a Time-Delay Neural Network Ensemble. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, pages 1–9.
- Trilla, A., Gratacòs, P., Guinart, D., Alessi, A. and Lamoureux, B. (2016). Health assessment of traction-motor blowers regarding their deformation degradation. *Proc. of the Third European Conference of the Prognostics and Health Management Society*, pages 712–718.
- Trilla, A., Janjua, F., and Bermejo, S. (2019). Developing a Hybrid Expert/Data-Driven Health Index for Railway Axleboxes Using Auto-encoder Neural Networks. *Proc. of the Prognostics and System Health Management Conference*, pages 1–6.
- Trilla, A., Mijatovic, N., and Vilasis-Cardona, X. (2022). Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting. *International Journal of Prognostics and Health Management*, 13(2):1–17.
- Trilla, A., Mijatovic, N., and Vilasis-Cardona, X. (2023). Unsupervised Probabilistic Anomaly Detection over Nominal Subsystem Events through a Hierarchical Variational Autoencoder. *International Journal of Prognostics and Health Management*, 14(1):1–15.
- Trilla, A., Miralles, D., and Fernández, V. (2020). Pushing Distributed Vibration Analysis to the Edge with a Low-Resolution Companding Au-

- toencoder: Industrial IoT for PHM. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, pages 1–8.
- Tsui, K. L., Chen, N., Zhou, Q., Hai, Y., and Wang, W. (2015). Prognostics and Health Management: A Review on Data Driven Approaches. *Mathematical Problems in Engineering*, 2015(793161):1–18.
- Turgis, F., Audier, P., Nemoz, V., and Marion, R. (2022). Health state characterization using clustering algorithms for railway fleet maintenance. *Proc. of the World Congress on Railway Research*, pages 1–6.
- UITP (2020). Digitalisation in Public Transport: Implementing Predictive Asset Maintenance. *Knowledge Brief*, pages 1–4.
- Unife (2017). A digital manifesto for Europe’s railways. *International Railway Journal*, 57(10):40–46.
- Unwin, D., and Sanzogni, L. (2022). Railway cyber safety: An intelligent threat perspective. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 236(1):26–34.
- Valcamonico, D., Baraldi, P., Amigoni, F., and Zio, E. (2022). A framework based on Natural Language Processing and Machine Learning for the classification of the severity of road accidents from reports. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 0(0):1–15.
- Valcamonico, D., Baraldi, Zio, E., Decarli, L., Crivellari, A., and La Rosa, L. (2024). Combining natural language processing and bayesian networks for the probabilistic estimation of the severity of process safety events in hydrocarbon production assets. *Reliability Engineering & System Safety*, 241(109638).
- van der Bijl, R., Utsunomiya, K., and van Oort, N. (2020). Failed projects offer valuable lessons for future schemes. *International Railway Journal*, 60(2):34–37.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. M., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *Proc. of the 31st Conference on Neural Information Processing Systems*, pages 1–15.
- Verdun, C., Audier, P., and Turgis, F. (2020). Digital transformation improves SNCF’s maintenance systems. *International Railway Journal*, 60(6):32–34.

- Vickerstaff, A., Bevan, A., and Boyacioglu, P. (2020). Predictive Wheel-rail Management in London Underground: Validation and Verification. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 234(4):393–404.
- Villani, C. (2018). For a Meaningful Artificial Intelligence. Towards a French and European Strategy. Technical report, French Digital Council.
- Villareal, J. F. (2019). The Hidden Front Line of Cyberattacks: Industrial Control Systems. *Uptime*, pages 42–44.
- Vollert, S., Atzmueller, M., and Theissler, A. (2021). Interpretable Machine Learning: A brief survey from the predictive maintenance perspective. *Proc. of the 26th IEEE International Conference on Emerging Technologies and Factory Automation*, pages 1–8.
- Wagstaff, K. L. (2012). Machine Learning that Matters. *Proc. of the 29th International Conference on Machine Learning*, pages 1–6.
- Wang, H., and Raj, B. (2017). On the Origin of Deep Learning. *arXiv:1702.07800 [cs.LG]*, pages 1–72.
- Wang, J., Yang, J., Bai, Y., Zhao, Y., He, Y., and Yao, D. (2021). A comparative study of the vibration characteristics of railway vehicle axlebox bearings with inner/outer race faults. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 235(8):1035–1047.
- Wang, Z., Mo, J., Gebreyohanes, M. Y., Wang, K., Wang, J., and Zhou, Z. (2022). Dynamic response analysis of the brake disc of a high-speed train with wheel flats. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 236(5):593–605.
- Ward, C., Goodall, R., Harrison, T., and Midgley, W. (2022). Mechatronic Applications in Rail Systems and Technologies. *EcoMechatronics*, pages 155–175.
- Wesendrup, K., and Hellingrath, B. (2020). A Process-based Review of Post-Prognostics Decision-Making. *Proc. of the European Conference of the Prognostics and Health Management Society*, pages 1–12.
- Wu, G., Dong, K., Xu, Z. *et al.* (2022). Pantograph–catenary electrical contact system of high-speed railways: recent progress, challenges, and outlooks. *Railway Engineering Science*, 30:437–467.

- Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. (2021). The Causal-Neural Connection: Expressiveness, Learnability, and Inference. *Proc. of the 35th Conference on Neural Information Processing Systems*, pages 1–54.
- Xin, T., Roberts, C., Weston, P., and Stewart, E. (2020). Condition monitoring of railway pantographs to achieve fault detection and fault diagnosis. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 234(3):283–300.
- Xu, M., Baraldi, P., Lu, X., and Zio, E. (2022). Generative Adversarial Networks With AdaBoost Ensemble Learning for Anomaly Detection in High-Speed Train Automatic Doors. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):23408–23421.
- Zanelli, F., Sabbioni, E., Carnevale, M., *et al.* (2023). Wireless sensor nodes for freight trains condition monitoring based on geo-localized vibration measurements. *Proc. of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 237(2):193–204.
- Zečević, M., Dhami, D. S., Veličković, P., and Kersting, K. (2021). Relating Graph Neural Networks to Structural Causal Models. *arXiv:2109.04173 [cs.LG]*, pages 1–29.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *Proc. of the International Conference on Learning Representations*, pages 1–15.
- Zhang, K., Huang, W., Hou, X., Xu, J., Su, R., and Xu, H. (2021). A Fault Diagnosis and Visualization Method for High-Speed Train Based on Edge and Cloud Collaboration. *Applied Sciences*, 11(1251):1–16.
- Zhang, K., Sun, Q., Zhao, C., and Tang, Y. (2023). Causal reasoning in typical computer vision tasks. *arXiv:2307.13992 [cs.CV]*, pages 1–17.
- Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., and Wei, M. (2019). A Review on Deep Learning Applications in Prognostics and Health Management. *IEEE Access*, 7:162415–162438.
- Zhang, P., Yang, Z., Moraal, J., Dollevoet, R., Zoeteman, A., and Li, Z. (2022). Laboratory investigation of effects of a friction modifier on wheel-rail dynamic contact. *Proc. of the World Congress on Railway Research*, pages 1–6.

- Zhu, J., Nostrand, T., Spiegel, C., and Morton, B. (2014). Survey of Condition Indicators for Condition Monitoring Systems. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, pages 1–13.
- Zoph, B., and Le, Q. (2017). Neural Architecture Search with Reinforcement Learning. *Proc. of the International Conference on Learning Representations*, pages 1–16.

INDEX

accelerometer, 23, 25, 28
Adam optimizer, 40
adhesion, 22
Adversarial, 40
aging, 17
Alstom, 6
Architecture Search, 41
Artificial Intelligence, 3
associational, 146
attention, 39
Autoencoder, 35
Autonomous driving, 18
axle, 24

battery, 21
Bayesian Network, 147
bias-variance, 42
bidirectionality, 41
Black-box model, 144
braking, 24
business case, 4
business model, 13

carbon footprint, 146
carbon strip, 26
catenary, 26
Causal Model, 146
CO₂, 21

compressor, 24
Computer Vision, 29
consensus, 145
context, 145
Convolutional, 37
cost-benefit, 16
counterfactual, 146
Cryptography, 20

data-driven, 33
decarbonizing, 21
decision-making, 17, 145
Deep Learning, 34
Denial-of-Service, 19
dense vector space, 39
detection, 43
diagnosis, 4, 43
digital twin, 33
distributed representation, 38
Double Descent, 42
Dropout, 37
dwell time, 18

edge computing, 28
electrification, 21
embedding, 36, 39
ensemble, 145
event data, 31

- explainability, 144
- failure, 16
- feature extraction, 37
- friction, 22, 27
- game theory, 150
- generative model, 40
- grease, 24
- health management, 43
- human resources, 17
- hydrogen, 21
- hypothesis, 8
- impact, 14
- Initialization, 37
- innovation, 3
- insights, 18
- inspection, 14
- Internet-of-Things, 28
- interpretability, 144
- interventional, 146
- jobs, 17
- knowledge distillation, 43
- leakage, 24
- local pattern, 37
- loss function, 40
- Lottery Ticket, 42
- maintenance, 44
- malware, 20
- management, 47
- mathematical analysis, 43
- Mobility, 18
- Multilayer Perceptron, 34
- neural networks, 34
- nominal variable, 31
- normalizing, 40
- operating condition, 19
- optical, 22
- overfitting, 37
- overhaul, 20
- overhead line, 27
- pantograph, 26
- parallelization, 42
- passenger flow, 18
- planning, 150
- pre-training, 36
- predictive maintenance, 4
- prognosis, 4, 43
- protection, 19
- railway, 4
- reconstruct, 36
- Rectified, 37
- recurrent, 39
- reduction, 15
- Reinforcement Learning, 39
- reliability, 16, 27
- Remote, 28
- Residual, 40
- reward, 39
- Root Cause, 147
- rules, 17
- safety, 4, 30
- scheduling, 150
- security, 19
- shortcut connection, 40
- sliding contact, 26
- supervised, 34
- sustainable, 20
- technical debt, 146
- Technical Language, 31
- temperature, 25
- text mining, 30
- threshold, 17
- topic model, 31

track, 23
training, 18
Transfer Learning, 41
Transformer, 41
transport, 3
travel pattern, 18

uncertainty, 145, 147
unsupervised, 34

value chain, 14
vanishing gradient, 37
Vibration, 25
vibration, 22, 24, 28

wear, 17
Wear rate, 22
wheel-rail, 21
work order, 31

