Explainable Machine Learning Models in Rail Track Maintenance

Muhammad Chenariyan Nakhaee[1], Djoerd Hiemstra[1] , Mariëlle Stoelinga[2,3]

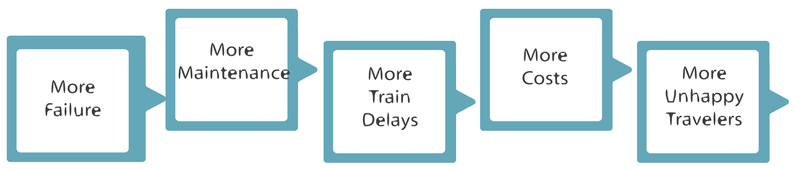
[1] Data Science, University of Twente, the Netherlands.

[2] Formal Methods and Tools, University of Twente, the Netherlands.

[3] Software Science, Radboud University Nijmegen, The Netherlands.

1. Introduction

The railway system in the Netherlands is the most important means of transportation. However, rails degrade over time which means:



But we can find a way to predict and prevent these failures and delays

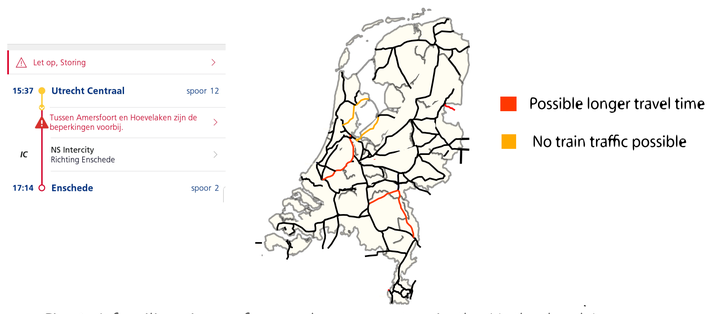


Fig. 1. A familiar picture for regular commuters in the Netherlands!

2. Machine Learning for Rail Maintenance

With the help of machine learning (ML) we can detect and predict defects to:

Efficiency ↑ Availability # of disruptions ↓ Costs

Two types of ML models have been used in the rail maintenace lliterature based on different data sources:

- Signal data ---> Shallow learning (e.g. random forests)
- Image data ---> Deep learning (e.g. convolutional neural networks)

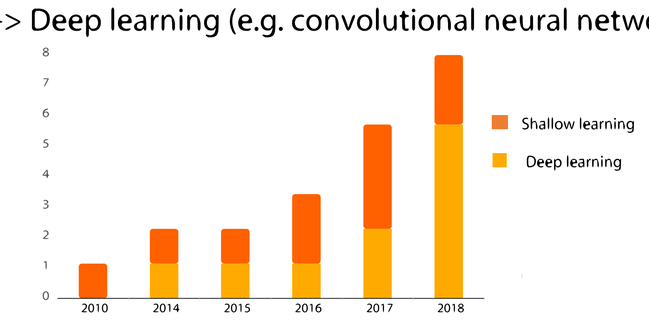


Fig. 2. increasing number of ML papers for rail track maintenance

3. In Deep Learning We Trust!

However, the majority of reviewed papers used black-box ML models such as CNNs for defect detection in rail tracks which are:

- Inherently non-interpretable
- Optimizied for a single metric such as classification accuracy.

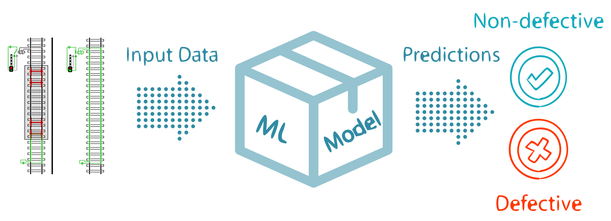


Fig. 3. Most of the ML models used in literature are black-box algorithms

In other words, an ML researcher might not able to fully explain how and why her CNN model came up with its predictions or prove its trustworthiness to the end user.

4. Explaining Machine Learning Models

Methods such as LIME can explain how and why an ML model makes a prediction.Yet, these techniques have their own limitations:

1. Designed for the researchers and not the end user
2. No consensus on how the explanations should be evaluated

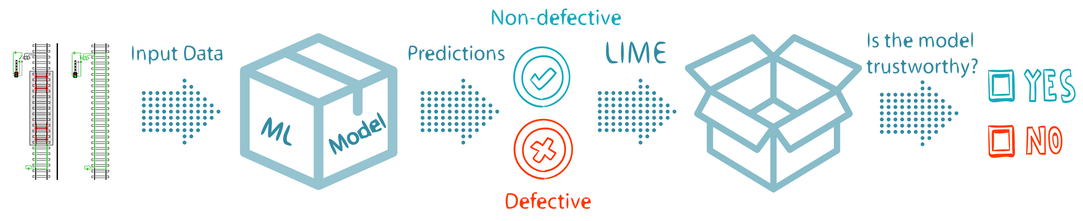


Fig. 4. Explaining a black-box model with LIME is possible

To overcome these limitations we propose two ideas:

- Evaluating the explanations with fault trees
- Using randomness to better explain an ML model

5. Evaluating Explanations with Fault Trees

Fault trees (FTs) are a widely-used and inherently explainable tool for safety and reliability analysis.

Data generated from FTs in FFORT repository can provide a baseline to better understand, evaluate and quantify the performance of existing techniques that provide explanations.

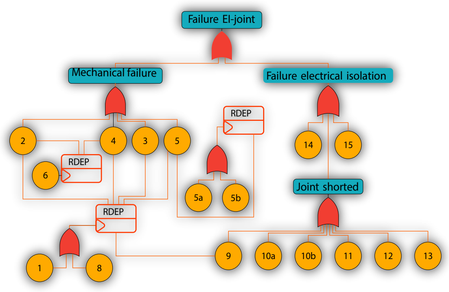


Fig. 5. Data generated from fault trees can be used to evaluate explainable ML techniques

FFORT (the Fault tree FOResT) which contains a collection of fault trees, collected from scientific literature is designed to provide such benchmark for the academic community.

6. Randomness Gives Better Explanations!

Intuitively, meaninngful explanations are better than the random explanations.

Therefore, by adding random features to our dataset, we can filter out the irrelavant and unimportant explanations.

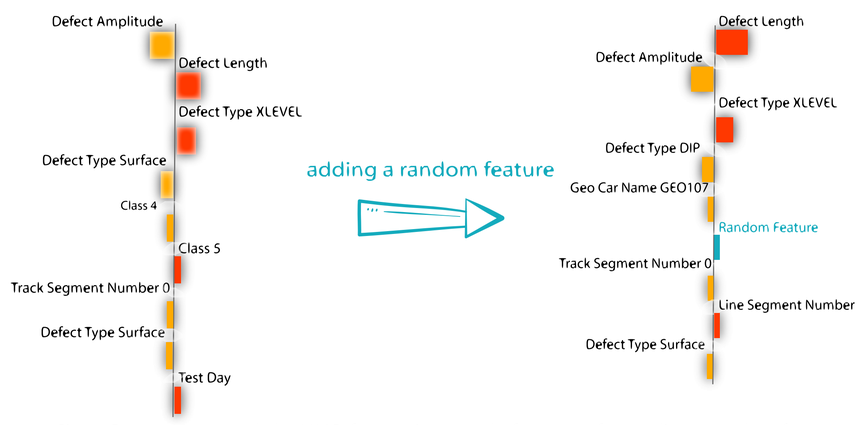


Fig. 6. Explanations made by LIME for one instance before and after adding a random feature using 2015 RAS Problem Solving Competition dataset

FFORT Benchmark

References

- Chenariyan Nakhaee M., Hiemstra D., Stoelinga M.I.A., van Noort M.: The Recent Applications of Machine Learning in Rail Track Maintenance: A Survey, RSSRail , 2019 (Accepted)
- Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. "why should i trust you?": Explaining the predictions of any classifier. KDD, 2016.
- Leilani H. Gilpin, Cecilia Testart, Nathaniel Fruchter, Julius Adebayo. "Explaining Explanations to Society". NeurIPS 2018 Workshop on Ethical, Social and Governance Issues in AI, 2018.

Acknowledgment

This research is supported by Prorail and the Netherlands Organization for Scientific Research (NWO) under the Sequoia project.

Netherlands Organisation for Scientific Research

UNIVERSITY OF TWENTE.

Commit2Data