AI Public Datasets for Railway Applications

Subjects: Engineering, Industrial | Computer Science, Artificial Intelligence

Contributor: Francesco Flammini, Mauro Pappaterra

The aim of this entry is to review existing publicly available and open artificial intelligence (AI) oriented datasets in different domains and subdomains of the railway sector. The contribution of this paper is an overview of AI-oriented railway data published under Creative Commons (CC) or any other copyright type that entails public availability and freedom of use. These data are of great value for open research and publications related to the application of AI in the railway sector.

Keywords: railways ; public datasets ; intelligent transportation ; machine learning ; predictive maintenance

1. Introduction

The automation of traditional manufacturing and industrial practices, by using modern technology in conjunction with massive recollection of data and powerful algorithms, has initiated the course of the fourth industrial revolution. Experts agree that AI is becoming the most central player in Industry 4.0, and the railway industry is not exempt from this. There are many applications within the railway sector in which AI can create a big impact [1]. The increasing number of IoT devices, data available and amount of computer power (along with the decreasing manufacturing costs for technology) create the conditions for the application of modern AI techniques in the railway sector [2].

Applications for AI models are diverse and can be implemented directly in vehicles, infrastructure and services related to transportation [3]. The recollection of data and use of AI algorithms can aid the analysis of data related to travel routes, the behavior of pedestrians and commuters, the mitigation of energy use, pollution, traffic congestion as well as the improvement of the overall security and safety of passengers. The European Union has assigned over €2.3 billion in funding for the development of smart, green and integrated transport for the period 2014–2020 within the Horizon 2020 Initiative [4]. This initiative comprises different research projects related to the application of AI in transport systems, including the Shift2Rails program, for the development and validation of sustainable, cost-efficient, high-performing, time-driven, digital and competitive train operation standards through railway research and innovation [5].

We recognize several literature reviews on specific AI applications in different railway subdomains. For instance, authors assessed literature on applying Machine Learning to track maintenance [6] and wheel defects [7], and [8] investigated image processing approaches for track inspection. Also, [9] addressed urban flow prediction using machine learning. For traffic management, [10] reviewed data-driven approaches. Similarly, [11] and [12] delved into big data for intelligent transport systems and railway systems, respectively, and [13] looked into potentials of swarm optimization for railways. Differently, [14] reviewed current AI developments across the railway sector holistically, covering all the above mentioned topics as well as safety and security, mobility, and autonomous train driving. They identified that majority of research belongs to maintenance and inspection (57%) and traffic planning and management (25%). The existing literature reviews typically covered a limited scope either regarding specific railway subdomains or some certain aspects of AI, with the exception of [14]. Also, all review papers addressed dominantly AI applications, with little/no focus on the used data. Here, we want to stress that the existence of available data is one of the critical aspects for AI applications. However, authors of AI-focused railway applications rarely publish the related data. In addition, railway companies still tend to be rather conservative and not open to sharing publicly their data to third parties, these being research and academic institutions. Therefore, relevant data are still rather scarce and often one of the largest challenges to address at the beginning of any research effort; according to the findings of the recent survey [15] the lack of suitable datasets for training the ML models has been indicated by the railway stakeholders among the top three obstacles to be faced for the adoption of AI in the rail sector. This situation altogether may lead to the delayed development of AI applications or even prevent researchers from starting to investigate specific topics.

The aim of this paper is to review existing publicly available and open datasets across the whole domain of railway sector for diverse AI applications. We cover subdomains such as maintenance and inspection, traffic planning and management, automated train driving, safety and security, passenger mobility and transport policy. Also, datasets are classified based on its type to numerical, image, label and other. The review covers the period until January 2021. We also present a short overview of supporting public datasets and APIs (Application programming interface (API) is a set of definitions and protocols for building and integrating application software which allows a product or service to communicate with other products and services). All these AI-oriented datasets are of great value for open research and publications related to the application of AI in the railway sector. With this study, we hope to benefit researchers in the fields of computer science and the transport industry by providing an insight into these valuable data and valuable information on how they can be accessed. In addition, companies would hopefully recognize the added benefit and be encouraged to share and publish their data more willingly.

2. Summary of the results

Based on previous literature reviews, we have analyzed publicly available datasets across seven railway subdomains for AI applications. We have used multiple portals to collect the public datasets including general databases like European Data Portal and Data.gov, AI-focused databases like Zenodo and FigShare and recent data-focused journals like Data in Brief. The data types have been classified as numeric, image, label and other.

We believe that with the public data available today, some railway problems could already be approached with AI-oriented solutions. For instance, the domain of Traffic Planning and Management counts with a good number of public datasets, in total 28 dataset, and vast availability of data harvesting sources, such as APIs, GTFS, RSS Feeds, for data collection. This data can be used for developing traffic predictions, and also timetabling and real-time rescheduling models and approaches. There are also rather complete and readily available datasets that can be used to research problems in the domain of Maintenance and Inspection, in total 28 datasets, which are mainly related to maintenance and inspection of railway tracks. Maintenance data for rolling stock, railway ballast and catenary system and electrical equipment are also present but to a lesser extent. This would be used for better health monitoring and predicting failures of infrastructure and rolling stock to minimize the disruption impacts on railway traffic. Other railway domains do not present as many openly available datasets: Safety and Security includes 10 public datasets and Passenger Mobility - 8 public datasets. Finally, no datasets could be found for Autonomous Train Driving, Transport Policy or Revenue Management. Thus, generating the first related public datasets could contribute greatly to boosting research in these domains. In all railway domains analyzed in this paper, the most common type of data is numerical data. We believe this is the easiest type of data to obtain, we have observed different data collection methods including mostly onboard sensors, and other IoT devices such as wireless sensor networks. However, only seven image datasets could be found. This is a counter-proportional considering the amount of research done involving computer vison models in the railway sector. Also, regarding the datasets quality, it often tend to be rather unknown; and limited or no proper documentation is sometimes available. On the positive side, the study revealed that there are publicly available data maintained by government organizations and TOCs that can be of great use to support AI-based models such as infrastructure network and statistical data, and certified APIs. In particular, these data coming from official public services and are usually well-curated real-time data that can greatly contribute to the accuracy of AI models. Moreover, the increasing number and sophistication of IoT devices – along with the decreasing manufacturing costs – present a promising outlook for the collection of raw data in the railway sector. There are different AI techniques related to Natural Language Processing (NLP) that could be utilized to process any unstructured data collected.

We recognize several promising research directions. First, the formation of a unified database of AI-oriented railway data would be beneficial, we believe that a centralized database of railway-specific datasets would greatly contribute to the conducting research in this area. Second, the further investigation of the quality of the existing datasets would be required to understand its size, quality, and applicability in greater details. Third, the collection of new high-quality data that can be made available for public use and research by active researchers and data and problem owners. Lastly, to guarantee the quality of new datasets, publications in peer-reviewed journals like Data in Brief and IEEE DataPort, and also online databases like Zenodo and FigShare shall be encouraged. We believe that these new developments would lead towards faster uptake and more diverse developments of AI applications in railway systems.

Read the full review and complete list of datasets: https://doi.org/10.3390/infrastructures6100136

3. Funding and disclaimer

This project has received funding from the Shift2Rail Joint Undertaking under the European Union’s Horizon 2020 research and innovation programme under grant agreement n. 881782 Rails. The JU receives support from the European Union’s Horizon 2020 research and innovation program and the Shift2Rail JU members other than the Union.

The information and views set out in this document are those of the author(s) and do not necessarily reflect the official opinion of Shift2Rail Joint Undertaking. The JU does not guarantee the accuracy of the data included in this document. Neither the JU nor any person acting on the JU's behalf may be held responsible for the use which may be made of the information contained therein.

References

1. Schwab, K. Foreign Affairs, The Fourth Industrial Revolution, What It Means and How to Respond. December 2019. Available online: https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution (accessed on 3 February 2020).
2. David, B. The Future of Intelligence is Artificial. International Railway Journal (IRJ). Available online: https://www.railjournal.com/in_depth/future-intelligence-artificial (accessed on 3 February 2020).
3. European Parliamentary Research Service (EPRS), European Parliament. Artificial Intelligence in Transport. Current and Future Developments, Opportunities and Challenges. April 2019. Available online: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRS_BRI(2019)635609_EN.pdf (accessed on 3 February 2020).
4. Innovation and Networks Executive Agency (INEA). Horizon 2020 Funding Areas. European Commission. Available online: https://ec.europa.eu/inea/en/horizon-2020 (accessed on 3 February 2020).
5. Shift2rail.org, "About". Available online: https://shift2rail.org/about-shift2rail/ (accessed on 3 February 2020).
6. Nakhaee, M.C.; Hiemstra, D.; Stoelinga, M.; van Noort, M. The Recent Applications of Machine Learning in Rail Track Maintenance: A Survey. In Proceedings of the International Conference on Reliability, Safety, and Security of Railway Systems, Lille, France, 4–6 June 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 91–105.
7. Thilagavathy, N.; Harene, J.; Sherine, M.; Shanmugasundari, T. Survey on railway wheel defect detection using machine learning. AutAut Res. J. 2020, 11, 4.
8. Liu, S.; Wang, Q.; Luo, Y. A review of applications of visual inspection technology based on image processing in the railway industry. Transp. Saf. Environ. 2019, 1, 185–204.
9. Xie, P.; Li, T.; Liu, J.; Du, S.; Yang, X.; Zhang, J. Urban flow prediction from spatiotemporal data using machine learning: A survey. Inf. Fusion 2020, 59, 1–12.
10. Wen, C.; Huang, P.; Li, Z.; Lessan, J.; Fu, L.; Jiang, C.; Xu, X. Train Dispatching Management With Data-Driven Approaches: A Comprehensive Review and Appraisal. IEEE Access 2019, 7, 114547–114571.
11. Zhu, L.; Yu, F.R.; Wang, Y.; Ning, B.; Tang, T. Big Data Analytics in Intelligent Transportation Systems: A Survey. IEEE Trans. Intell. Transp. Syst. 2018, 20, 383–398.
12. Ghofrani, F.; He, Q.; Goverde, R.; Liu, X. Recent applications of big data analytics in railway transportation systems: A survey. Transp. Res. Part C Emerg. Technol. 2018, 90, 226–246.
13. Wu, Q.; Cole, C.; McSweeney, T. Applications of particle swarm optimization in the railway domain. Int. J. Rail Transp. 2016, 4, 167–190.
14. Bešinović, N. Deliverable D1.2: Summary of Existing Relevant Projects and State-of-the-Art of AI Application in Railways, RAILS, Shift2Rail. Available online: https://rails-project.eu/wp-content/uploads/sites/73/2021/05/RAILS_D12_v23.pdf (accessed on 5 April 2020).
15. Marrone, S.; De Donato, L.; Vittorini, V.; Nardone, R.; Tang, R.; Besinovic, N.; Flammini, F.; Goverde, R.M.P.; Lin, Z. Findings about the State-of-Practice. Deliverable D1.3 Application Areas (Chapter 5). 2021. Available online: https://rails-project.eu/wp-content/uploads/sites/73/2021/09/RAILS_D1_3_Application_Areas_v32.pdf (accessed on 15 August 2021).