Artificial intelligence in railways

Current applications, challenges, and ongoing research

De Donato, Lorenzo; Tang, Ruifan; Besĭnović, Nikola; Flammini, Francesco; Goverde, Rob M.P.; Lin, Zhiyuan; Liu, Ronghui; Marrone, Stefano; Napoletano, Elena; More Authors

DOI
10.4337/9781803929545.00017

Publication date
2023

Document Version
Final published version

Published in
Handbook on Artificial Intelligence and Transport

Citation (APA)
De Donato, L., Tang, R., Besĭnović, N., Flammini, F., Goverde, R. M. P., Lin, Z., Liu, R., Marrone, S., Napoletano, E., & More Authors (2023). Artificial intelligence in railways: Current applications, challenges, and ongoing research. In H. Dia (Ed.), *Handbook on Artificial Intelligence and Transport* (pp. 249-283). Edward Elgar Publishing. https://doi.org/10.4337/9781803929545.00017

9. Artificial intelligence in railways: current applications, challenges, and ongoing research

Lorenzo De Donato, Ruifan Tang, Nikola Bešinović, Francesco Flammini, Rob M.P. Goverde, Zhiyuan Lin, Ronghui Liu, Stefano Marrone, Elena Napoletano, Roberto Nardone, Stefania Santini, and Valeria Vittorini

I. INTRODUCTION

Artificial intelligence (AI) is increasingly being affirmed as a game-changer technology in several sectors, including transport. Despite automotive and avionics being the most explored fields, there is growing interest in the application of AI to railway systems, both as a key factor and in conjunction with other prominent technologies such as cloud computing, big data analytics, and the Internet of Things (IoT). This trend, further supported by the achievements obtained in other relevant transport sectors, is also witnessed by several industrial research and innovation initiatives as well as by the growing number of scientific studies focusing on the use of AI in the rail sector. AI is expected to significantly impact many railway areas, such as autonomous and cooperative driving, predictive maintenance, and traffic management optimization, in a medium to long-term perspective, benefitting by increasing line capacity, reducing life cycle cost, improving human error detection and avoidance, and enhancing efficiency and performance. Despite its potential in opening unprecedented scenarios, AI also poses significant challenges and raises several concerns, spanning from ethics to trustworthiness.

The uncertainty and instability of specific AI approaches do not allow the certifications of AI systems against current safety standards. Therefore, besides the potential of AI as a means, some regulatory issues arise when it comes to allowing AI systems to directly control trains or railway assets.

This chapter presents and discusses some of the results of the H2020 Shift2Rail (S2R) project RAILS – Roadmaps for AI integration in the rail sector (RAILS 2019) – and further ongoing research. The main objective is to provide a bird's eye view of the main challenges and opportunities for AI in railways and related issues and regulatory concerns based on past and ongoing research in the rail sector. To that aim, Section II provides a high-level overview of AI applications worldwide introducing preliminary challenges. Section III analyses the results of a survey that involved different stakeholders between railway organisations and research centres. Section III then discusses problems concerning the physical application of AI in real-world scenarios and the challenges encountered when designing, developing, and testing some of these applications with particular emphasis on data-related issues. Section IV summarises the efforts made so far towards the definition of standards and regulations to drive the realisation of ethical, reliable, and trustworthy AI systems with a particular emphasis on European directives. Some of these regulations might not be perfectly tailored to the rail sector and might not delineate practical procedures, however, they introduce several valuable

concepts and criteria that could efficiently drive the design of AI systems in railways. Section V discusses some of the current open railway problems that could benefit from AI and highlights some preliminary solutions and directions that could support future research on AI in railways. Lastly, Section VI provides some closing remarks.

II. CURRENT AND FUTURE AI APPLICATIONS IN RAILWAYS

Many different definitions and meanings of "artificial intelligence" have been introduced, starting with Turing's test, whereby "a machine is deemed intelligent if it is indistinguishable from a human during a conversation with an impartial observer" (Turing 2009). More recently, the European Commission provided this definition: "Artificial intelligence refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals" (European Commission 2018). However, to highlight the potential of AI in railways, a more tailored definition is needed (Bešinović et al. 2022) as follows:

> the discipline gathering all the aspects that allow an entity to determine how to perform a task and/or take a decision based on the experience matured by observing samples and/or by interacting with an environment, possibly competing against or cooperating with other entities.

Bešinović et al. (2022) also present a taxonomy of AI and related concepts to discern between fundamental AI aspects and concepts that are commonly attributed to AI but do not strictly belong to it, and empirically subdivide the railway domain into seven railway areas to structure a preliminary overview of AI in railways.

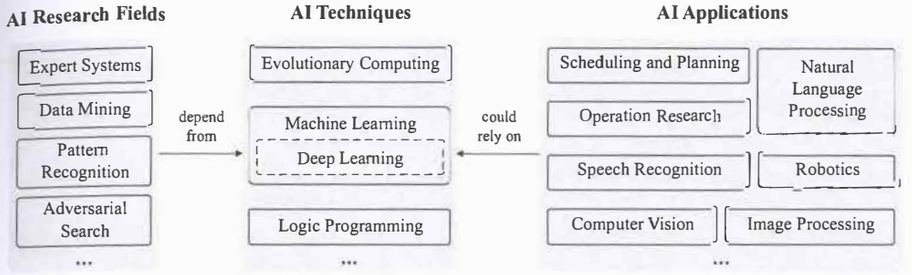
According to the taxonomy, AI is seen as an integrated concept encompassing:

- AI techniques: the set of approaches that enables entities to perform tasks commonly associated with intelligent behaviour (e.g., machine learning (ML), evolutionary algorithms).
- AI research fields: research areas that directly rely on AI techniques and would not exist without them (e.g., expert systems, pattern recognition).
- AI applications: areas that could benefit from AI techniques but are not an integral part of AI (e.g., computer vision (CV) operation research).

Figure 9.1 shows most of the AI concepts classified according to this taxonomy.

Concerning the rail sector, the authors identified seven railway areas (reported in Figure 9.2) based on the current literature with the aim of providing a preliminary overview of AI applications in railways. Then, based on the same taxonomy and railway areas, Tang et al. (2022) and RAILS (2021a) carried out a more detailed survey on the AI techniques exploited to address railway tasks. Particularly, Tang et al. (2022) mainly focused on scientific studies, while RAILS (2021a) also analysed projects conducted worldwide with particular emphasis on those developed within the European S2R framework. Other contributions can also be found within the literature addressing, whether horizontally or more vertically, the advancements of AI in railways. Some of these surveys and reviews are reported in Table 9.1.

The authors reported some of the main railway problems (summarised in Figure 9.3) that have been faced or are currently challenging researchers and practitioners according to Tang



Source: Authors.

Figure 9.1 *A taxonomy of artificial intelligence*

Maintenance and Inspection	Preventive or corrective activities intended to keep a railway system or subsystem in proper operating condition.
Traffic Planning and Management	Solutions oriented to an efficient capacity management, timetabling, control of railway operations, resource allocation, and resource management.
Safety and Security	Activities aiming at reducing the risks of both unintentional and intentional accidents that may hurt persons and/or damage physical assets.
Passenger Mobility	Applications oriented at analysing crowd flows and similar as to provide efficient solutions to move people through the railway network.
Autonomous Driving and Control	Solutions oriented at dynamically supervising train operations and/or making trains capable of taking autonomous driving decisions.
Transport Policy	Strategies and programmes to achieve specific objectives related to social, economic and environmental conditions, and railways' performance.
Revenue Management	Applications oriented at predicting consumer behaviour at the micro-market levels to optimise product availability and price to maximise revenue growth

Source: Authors.

Figure 9.2 *Railway areas*

et al. (2022) and RAILS (2021a). These manuscripts represent two of the most comprehensive reviews among those mentioned above and can help to provide a high-level overview of current AI integration in railways.

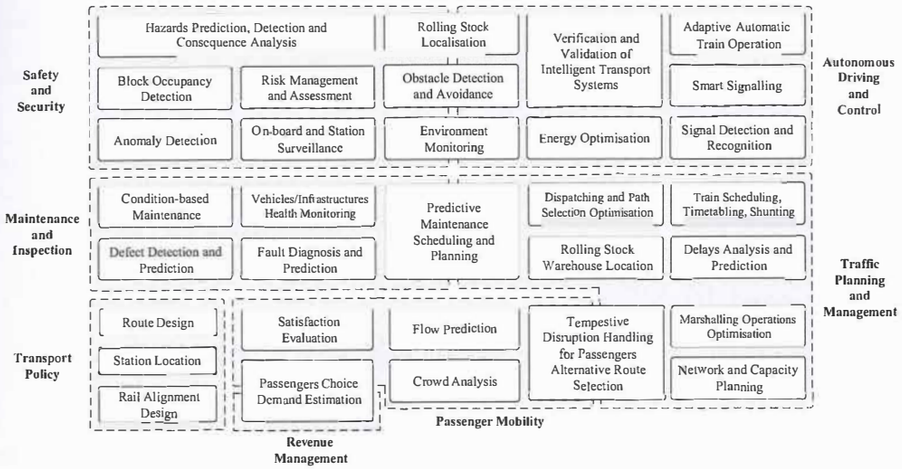
Maintenance and inspection. Several AI solutions have been proposed to deal with defect detection, fault diagnosis, fault and anomaly detection, failure prediction, maintenance planning (i.e., dynamic scheduling of maintenance and inspection activities) concerning infrastructures (e.g., tunnels, bridges), rail tracks, catenary systems, and rolling stock components. Almost any kind of AI technique has been investigated, starting from evolutionary computing approaches to traditional ML techniques (e.g., decision trees, random forests, support vector machines) and advanced CV solutions based on deep learning (DL) solutions. Particular

Table 9.1 *Some surveys and reviews addressing AI in railways*

Reference	Covered railway areas (or topics)	Main investigated AI fields
Beš inović et al. 2022	All railway areas indicated in Figure 9.2	Most of the AI-related concepts in Figure 9.1
De Donato et al. 2022	Maintenance and inspection	Image processing, computer vision, and audio processing
Pappaterra 2022	Maintenance and inspection	Majority of AI research fields and AI techniques from Figure 9.1
Tang et al. 2022	All railway areas indicated in Figure 9.2	Most of the AI-related concepts in Figure 9.1
Fayyaz et al. 2021	Obstacle detection at level crossings	Image processing and computer vision
RAILS 2021a	All railway areas indicated in Figure 9.2	Most of the AI-related concepts in Figure 9.1
Ristrić-Durrant et al. 2021a	Obstacle detection on rail tracks	Image processing and computer vision
Liu et al. 2019	Maintenance and inspection	Image processing and computer vision
Xie et al. 2020	Crowd, traffic, and transit flow prediction	Machine learning, reinforcement learning, deep learning
Yin et al. 2020	Maintenance and inspection, timetabling, dynamic price and seat allocation, and speed and trajectory control	Data mining, expert systems, genetic algorithms, machine learning, image processing, and computer vision
Chenariyan Nakhaee et al. 2019	Rail track maintenance and inspection	Machine learning, deep learning, image processing, and computer vision
Ghofrani et al. 2018	Maintenance and inspection, traffic planning and management, and accident analyses	Machine learning, deep learning, image processing, and computer vision (but also big data analytics approaches)

attention is recently being given to the continuous monitoring and fault prediction of some critical assets such as level crossings (discussed in Section V.A).

Traffic planning and management. Although several contributions have been found regarding the usage of AI techniques for traffic analysis, delay prediction, timetabling, train shunting in station areas, and real-time rescheduling, widespread AI integration is still in its infancy. In this context, different AI approaches have been investigated including ML-based regression models and other DL techniques including convolutional neural networks (CNNs) and recurrent neural networks to deal with traffic analysis and delay prediction. At the same time, the usage of deep reinforcement learning based on Q-learning or actor-critic algorithms has recently caught the attention of researchers to deal with scheduling and management problems. In addition, to overcome the lack of topology information and structural dependencies, graph embedding-based approaches can be adopted to interpret complex network features to make delay prediction processes more accurate.



Source: Authors.

Figure 9.3 *Investigated railway problems*

Safety and security. Investigations within this area were mainly oriented at analysing CV and DL approaches to detect threats within the environment (whether stations or rail tracks) that could affect the safety of railway personnel and customers. Attempts were also made towards the realisation of AI systems for the qualitative evaluation of risks on physical sites and the construction of an unsupervised learning approach to detect anomalies in railway systems. This application area presents some real challenges with regard to hazard prediction: the main problem is that events are so rare that it is extremely challenging to build a suitable dataset to properly train AI models.

Autonomous driving and control. The usage of CV and DL approaches to deal with obstacle detection on rail tracks is a cross-border application that also falls under this area. Other applications that have been discussed are signal recognition, where CV approaches may be beneficial, and the characterisation of energy-optimal driving profiles, as well as leveraging reinforcement learning approaches. Besides these specific applications, there are some issues related to certifications and the standardisation of AI approaches when it comes to safety-critical applications in general (to be discussed in the following sections) that are consistently slowing the process of integration of AI in this context.

Passenger mobility. In this context, deep learning has been found to be largely applicable and better performing than traditional ML algorithms when it comes to flow predictions. Also, investigations have been carried out towards crowd analysis, crowdedness prediction, traveller pattern recognition, and traveller clustering. From the perspective of application scenarios, the literature has a specific focus on passenger movement simulation and transfer navigation in stations.

Revenue management and transport policy. As for these two railway areas, limited contributions were found. For transport policy, the contributions only included interesting hints

about applications for which AI could be useful including route design, taking into account the geological conformation, and station location. The same goes for revenue management, which would benefit from newly available data and AI techniques for applications like ticket pricing prediction or seat booking control.

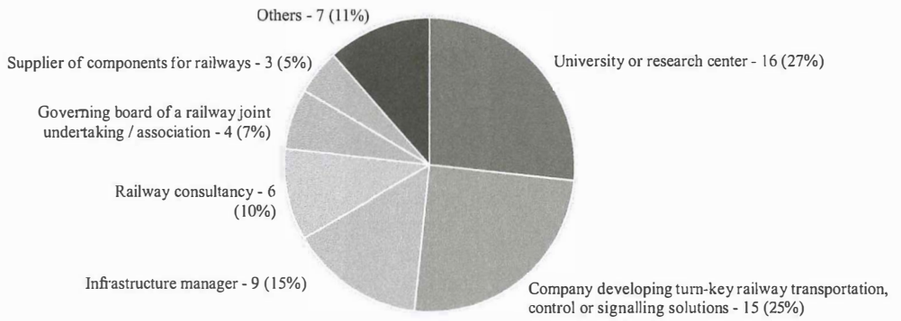
The in-depth analysis of these problems is out of the scope of this chapter. Instead, in the following sections, the main challenges that have been encountered when designing or implementing AI solutions will be presented and discussed, as well as the main issues that are slowing the process of integration of AI in railways, especially when it comes to safety-critical applications. Then, in Section V, some railway problems that are currently considered hot topics within the railway community will be analysed.

III. ISSUES AND CHALLENGES

From a theoretical perspective, most of the AI solutions proposed in the literature, especially those related to maintenance and inspection, can be considered mature enough. Besides the possibility for further optimization and improvements, it has been demonstrated that AI techniques (if properly selected) can effectively detect trends and changes in data. However, there are some challenges and issues that affect the effective and practical introduction of AI in railways that it is worth underlining to possibly delineate some directions for future improvements. The results from a survey conducted in the context of the RAILS project (RAILS 2021b) are presented that could help to identify the main issues to be overcome and some useful indications to consider for the effective adoption of AI applications in the railway sector.

The survey, conducted in 2020, involved about 60 participants from ten research institutes and 30 railway companies worldwide. Figures 9.4 and 9.5 respectively show the distribution of the participants according to the organisational typology and geographical region of their affiliation.

The survey included several types of questions related to the adoption of AI in railways. This section, however, mainly focuses on the challenges that should be overcome, the milestones that should be reached, and some dataset issues that should be addressed for the fast



Source: Authors.

Figure 9.4 Distribution of respondents according to their affiliation's typology

Source: Authors.

Figure 9.5 Geographical region of operation of respondents' organisations

take-up of AI in railways. These issues have been marked in the literature and project analyses as summarised in Section II. The reader is also referred to RAILS (2021b) for further details.

III.A Challenges, Potential Barriers, and Key Milestones

The survey presented to the participants some challenges in terms of potential barriers, that is, factors that are blocking the practical integration of AI in railways. The participants were asked to associate each of these barriers with a score ranging from 1 (not blocking) to 6 (extremely blocking). Table 9.2 ranks the barriers based on the average scores computed by considering the responses from all the participants. Notably, it also shows the rankings from research centres (including universities) and organisations separately because different entities may have different goals that could be affected by factors in different manners.

In the same way, the participants were asked to rank five milestones that could drive the integration of AI in railways by associating each of them with a score ranging from 1 (less relevant) to 6 (most relevant). Table 9.3 reports the results of the survey in relation to the key milestones.

To summarise, the "safety, dependability, and trustworthiness concerns" that affect AI systems were found to be the most blocking barrier according to the research community. This is clearly evident from Table 9.3 where the realisation of explainability and trustworthy AI approaches is seen as the most relevant milestone to be reached. Conversely, practitioners are more interested in data and knowledge sharing but also expressed concerns about the status of digitalisation of the current railway systems that might not be so ready to accommodate AI applications.

Interestingly, the respondents were also asked to rate the level of maturity of AI approaches in terms of practical potential for implementation in current railway systems. The results show that, after the digitalisation status of the railways, AI approaches oriented at smart/predictive maintenance are most likely to reach good maturity levels in the following few years. Conversely, those AI approaches that should directly be involved in safety-critical applications would take a longer time to be considered effective because of both the current shortcomings of the AI approaches themselves and the lack of ad-hoc standards and regulations. Overall, the maturity level of AI in railways achieved an average score of 3 (possible range is between 1 and 6), and nobody rated AI to be mature enough to be applicable in the short term.

Table 9.2 Ranking of barriers

Barrier	Research centre	Railway organisation	Average score
Safety, dependability, and trustworthiness concerns	5.00 (1)	4.16 (4)	4.39
Lack of datasets for testing	4.69 (2/3)	4.25 (1)	4.37
Missing specific standards or regulations	4.44 (5)	4.20 (2)	4.27
Insufficient level of digitalisation in railways	4.50 (4)	4.17 (3)	4.26
Lack of knowledge and competences about AI	4.13 (6)	4.05 (5)	4.07
Privacy and confidentiality concerns	4.69 (2/3)	3.84 (7)	4.07
Lack of high-level design principles and guidelines	4.00 (7)	3.86 (6)	3.90
Lack of means to share knowledge among stakeholders	3.94 (8)	3.63 (8)	3.71

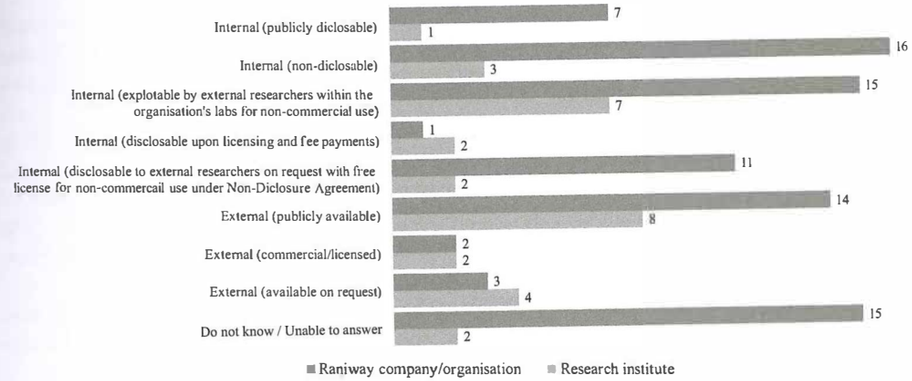
Table 9.3 Key milestones towards the Integration of AI

Key milestone	Research centre	Railway organisation	Average score
Creation of technical infrastructures at the European level for sharing data and knowledge (datasets, results, models, etc.)	4.80 (2)	4.56 (1)	4.62
Specific and in-depth study on the development and applicability of explainable and trustworthy AI approaches	4.81 (1)	4.48 (2)	4.57
Establishment of European working groups in railway organisations specifically addressing the cooperative design of AI approaches	4.73 (3/4)	4.27 (3)	4.39
Exploitation of cloud and edge computing, including IoT, at the railway network level to support data-driven approaches	4.38 (5)	4.25 (4)	4.28
Creation of transversal working groups on AI in critical systems to discuss, share, and learn from other sectors	4.73 (3/4)	4.12 (5)	4.28

III.B Data Issues

Concerning datasets, the participants were asked to indicate the typology of the datasets they were using/working on. Their responses are summarised in Figure 9.6.

Considering that each respondent could select multiple responses (if, for example, they were working on different tasks by exploiting different datasets), it is interesting to analyse the willingness of practitioners to disclose datasets that were built internally in their organisations. Provided that the thoughts of respondents do not directly reflect the willingness of their organisations, only seven practitioners considered their datasets to be publicly disclosable and



Source: Authors.

Figure 9.6 Typology of datasets used by railway organisations and research institutes

available for public use. The other participants did not disclose whether their datasets were shared with other researchers or if the research activities were only conducted within their organisations. Then, in some cases, they reported being available to disclose their data under non-disclosure agreements.

The fact remains that the effectiveness of some AI approaches and their safe implementation in real environments strongly relies on data. Therefore, a brief discussion on how to deal with the problem of data quality and availability is proposed below. The next section also provides suggestions about the current regulations and guidelines on the usage of AI, the ongoing standardisation effort, and related technical aspects such as explainability.

III.B.1 Data issues

AI models are generally capable of extracting knowledge from almost every kind of data. However, the performance of AI models could be heavily affected by data that do not reflect the real world or describe only a particular subset of real scenarios. In other words, data that do not express relevant information to tackle a given task will hamper model development. In this context, *data quality* is related to the procedure adopted to build the dataset; the kind of sensor used to collect data; and the class of attributes (e.g., temperature, vibration) that describe data. There are cases in which scientific studies (e.g., some of those discussed by De Donato et al. 2022) have addressed the same problems (e.g., rail track defect detection) but adopted different proprietary datasets built from scratch. Since the dataset is an integral part of an AI model, which means that by changing the dataset the proposed AI model may not perform as expected, these approaches may not be properly compared to each other and, more importantly, if the dataset is not built correctly, these approaches may solve problems related to a distorted version of reality. Hence, even though they are all theoretically suitable solutions, it is challenging to understand which approach is more effective than others. This problem becomes more sensitive when it comes to autonomous train driving. For example, assuming that a supervised vision-based DL algorithm is trained to recognise a set of obstacles on rail tracks (as most of those discussed by Ristić-Durrant et al. 2021a) and that "leaves"

are not considered in the training phase; what would happen if a bunch of leaves, which could reduce the adhesion between train's wheels and rails, were on rail tracks? Most likely, the system would not recognise them as it has not been trained to do that. In other words, it has been fed with only a subset of reality.

This is just one simple example highlighting the possible criticalities of AI systems. Most likely, obstacle detection systems would leverage data coming from different sensors (e.g., cameras, radars, and LiDARs) that could compensate each other to efficiently detect different types of obstacles (see, for example, SMART2 2019). Nevertheless, also in this case, if data collected to train AI modules do not reflect reality, it would not be possible to efficiently train AI models capable of operating effectively in real environments.

The data quality problem could be mitigated by introducing *standardised datasets* that could be used as benchmarks to train and test AI applications. Examples are ImageNet (Deng et al. 2009) and MS COCO (Lin et al. 2014), which have been extremely useful for CV applications. A "standardised" dataset is established by railway stakeholders and bodies who have great expertise on the task or the class of tasks for which the dataset is built and agree on the attributes, the samples, and the specific procedures to collect them. In this way, it would be possible to *develop suitable and easily comparable AI solutions*. In addition, or alternatively, a set of guidelines to properly build a good-quality dataset could be defined.

III.B.2 Data availability

Data availability is another sensitive issue that is affecting the development of AI solutions. Practically, the term "data availability" has two main connotations: (i) proprietary datasets are not shared with the community, therefore, it is not straightforward to reproduce some studies or propose alternative solutions based on the adopted data; and (ii) the events that should be analysed through AI models are rare by nature (e.g., crashes and safety-critical systems' failures), therefore, their data are extremely difficult to collect. Some techniques exist that help to build at least preliminary versions of AI systems.

Assuming that a small dataset is available, data augmentation approaches could help to substantially increase the dataset by generating synthetic data and building AI solutions that could work properly on real data. There are different data augmentation approaches: some of them are purely algorithmic, hence, they do not rely on AI. Other approaches are based on AI techniques, such as generative adversarial networks (Antoniou et al. 2017).

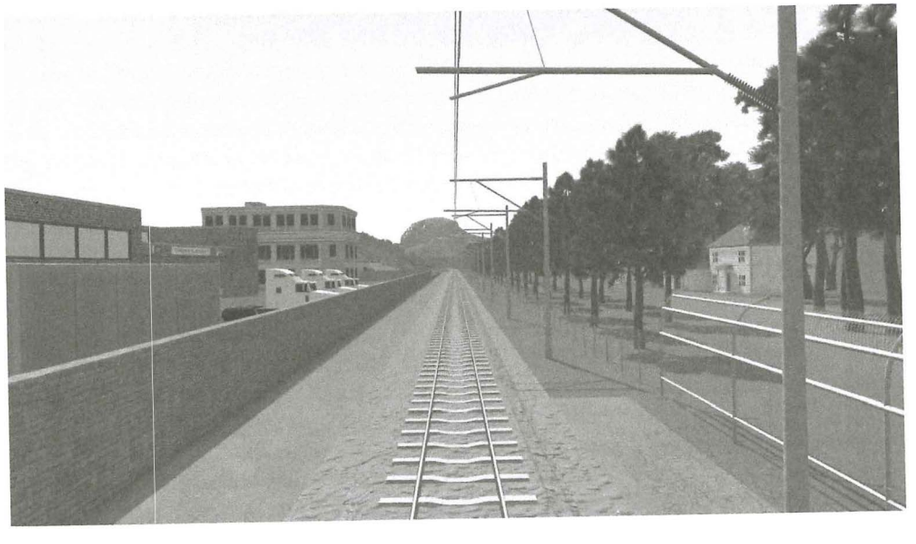
Alternatively, transfer learning (Pan and Yang 2010) approaches could be exploited. These aim at reusing the knowledge acquired to solve a given task (source task) in a specific application field (source domain) to address a similar or different problem (target task) in the same or another domain (target domain). Also in this case, it is possible to obtain suitable performance with a limited amount of target data (i.e., data related to the target task to be solved). However, data related to the source task (source data) must be available and, to some extent, source and target data must be related, otherwise, the knowledge extracted from the source data would not be properly usable to analyse target data.

Further, federated learning approaches (Yang et al. 2019) could also work in case data exist but owners are not intending to share them. Indeed, through federated learning, it would be possible to train an AI model on a dataset, then share the model (without the data) with other partners that will continue the training on their data and so on.

The previous solutions work on the assumption that an initial dataset is available. In any case, it would be useful to know the publicly available datasets in the railway domain. Pappaterra et al. (2021) performed a systematic literature review of railway datasets that are currently available online. About 62 datasets were found, most of which are related to "traffic planning and management" and "maintenance and inspection applications" railway areas. Only a few datasets were found in the context of autonomous train driving, in such a context, Cityscapes (Cordts et al. 2016) or RailSem19 (Zendel et al. 2019) could be exploited for semantic segmentation applications to allow trains to perceive their surrounding environments.

Other solutions for coping with the problem of data availability are quite independent of AI. They expect to leverage digital twins (DTs) or simulators, or a combination of both, to collect synthetic data. Simulators would be useful when it comes to collecting, for example, images or point cloud data that represent objects or space. Products that could be useful in this context can be subdivided into two macro-categories: videogames and 3D editors. Videogames (e.g., Grand Theft Auto 5) offer high graphic characteristics, close to reality, but the customisation of the environment is not so straightforward. On the other hand, 3D editor software like the MathWorks' RoadRunner – originally taught for automotive scenarios but adaptable to railways (see Figure 9.7 for an example) – offer a high customisation of the environment, but they could not provide the same graphic characteristics as videogames.

In addition to simulators, digital twins also may play a central role to collect data, especially in situations where the events being analysed are rare (e.g., failures of safety-critical systems). A digital twin can be seen as a virtual representation of a physical asset that is linked to and evolves with it by means of models and IoT sensors (Kritzinger et al. 2018). Hence, DT is nothing more than a digital model that has a physical counterpart and, once properly built, could represent the real system in the digital world. Therefore, it would be possible to adopt some techniques such as fault injection (Orive et al. 2019) to stimulate failures in order to



Source: Authors.

Figure 9.7 *A railway scene built in RoadRunner 3D Scene Editor*

safely record data related to malfunctions. The reader is referred to Dirnfeld et al. (2022) for a review of key railway applications of digital twins.

IV. STANDARDS, REGULATIONS, AND TRUSTWORTHY AI

The lack of properly developed standards for AI, together with ethical principles and safety, dependability, and trustworthiness aspects, also play a key role in slowing the process of integration of AI. The effort of standardising AI is progressing on the right path, but several practical difficulties complicate this process, in particular when AI has to be used for realising safety-critical functionalities.

However, it is important to mention that although all these factors are generally relevant in railways and other transport sectors, not all the applications pose safety-critical risks to justify legislative interventions (Besˇinović et al. 2022). For example, AI applications oriented towards identifying or predicting faulty components do not involve relevant safety-critical requirements. On the other hand, some ethical concerns could arise when applying AI to staff scheduling, where the wellness of human beings must be ensured. Finally, major ethical concerns arise when it comes to safety-critical applications including autonomous train driving. In this context, standards and practical regulations on the implementation and formal verification of AI systems are necessary.

IV.A A European Vision of Trustworthy AI

At the European level, the High-Level Expert Group on AI (AI HLEG) provided a definition for "trustworthy AI" and some general guidelines in terms of *ethical principles* (EP) and *key requirements* (KR) that AI systems should meet (AI HLEG 2019). Trustworthy AI is defined as an integrated concept according to which AI systems must be as follows: *lawful*, meaning that they must respect all applicable laws and regulations; *ethical*, meaning that they should follow ethical principles and values; and *robust*, referring both to the technical perspective and to the need of taking into account its social environment. Four ethical principles and seven key requirements that AI systems must meet to be deemed trustworthy were also outlined (Table 9.4).

From a high-level perspective, EPs establish that AI systems must neither subordinate human beings nor affect their rights or wellness. Then, AI systems must not limit users' freedom of choice and humans must have the possibility to contrast the decisions taken by AI, which intrinsically implies that AI systems must be comprehensible and transparent.

Table 9.4 AI HLEG's ethical principles and key requirements

Ethical principles	Key requirements
EP1: Respect for the freedom and autonomy of human beings	*KR1*: Human agency and oversight
EP2: Prevention of any kind of harm	*KR2*: Technical robustness and safety
EP3: Principle of fairness	*KR3*: Privacy and data governance
EP4: Principle of explainability	*KR4*: Transparency
	KR5: Diversity, non-discrimination, and fairness
	KR6: Societal and environmental well-being
	KR7: Accountability

Key requirements conceptually schematise what is stated through the EPs and provide high-level guidelines about AI system design and implementation. Analysing them from the railways' perspective: human operators must be able to supervise and accept or decline (human-in-the-loop/human-over-the-loop) an action taken or suggested by AI systems (KR1); AI systems must be resilient, safe, reliable, and reproducible as to mitigate potential harms (KR2); data privacy and protection as well as data quality and integrity must be ensured (KR3); AI systems must be understandable by all the possible stakeholders including developers, railway operators, and customers who should also be aware they are interacting with AI systems and must be informed about system's capabilities and limitations (KR4); AI systems' decisions must not be driven by biased or discriminatory reasoning, and should be accessible to everyone (KR5); AI systems must be designed and implemented to be useful and beneficial for as many individuals as possible; they must be environmentally friendly (KR6); and AI systems must be transparent, comprehensible, and auditable to allow for the reconstruction of the responsibility chain in case of malfunction or crashes (KR7).

It is worth highlighting here that such KRs have subsequently led to the definition of "the assessment list for trustworthy AI (ALTAI) for self-assessment" (AI HLEG 2020), a framework studied to qualitatively evaluate and, indeed, self-assess the trustworthiness of AI systems. The ALTAI framework encompasses, for each of the KRs, a list of criteria that should drive the design of AI systems. Although these criteria are more practical than the aforementioned KRs, they do not delineate any specific process that should be adopted when designing and implementing AI systems; they only establish some characteristics the systems should have. Pathways to implement and prove the compliance of the systems become the responsibility of developers.

In addition to these guidelines, it is worth highlighting that the European Commission established in the past years some strict regulations on data governance and allowed AI applications which affect (directly or indirectly) the introduction of AI in railways. The "General Data Protection Regulation (GDPR)" (European Commission 2016) delineates some strict regulations to guarantee that European citizens are the only owners of the data associated with their lives and activities. It also forces companies (European or not) to be compliant with the regulations when they manage European citizens' data. Given that AI systems are extremely dependent on data, GDPR is of crucial importance since data collection, processing, storage, and management should be done in a fully GDPR-compliant manner. Also, GDPR's Article 22 poses important limitations on the use of systems capable of taking autonomous decisions that can impact users. Along this line, the "Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)" (European Commission 2021) relies on and extends the aforementioned documents and introduces a cross-domain and horizontal regulatory framework oriented at preserving fundamental human rights (RAILS 2021b). Through this document, the European Commission banned all AI applications that may violate fundamental rights or could manipulate persons beyond their consciousness in order to distort their behaviour. For example, AI-based social scoring applications are prohibited. Then, the framework collects a set of requirements that high-risk AI systems, including "AI systems intended to be used as safety components in the management and operation of road traffic", must meet in order to be marketed and operated in Europe. These include documentation, tests, risk assessment, data quality evaluation, and event logging. Although

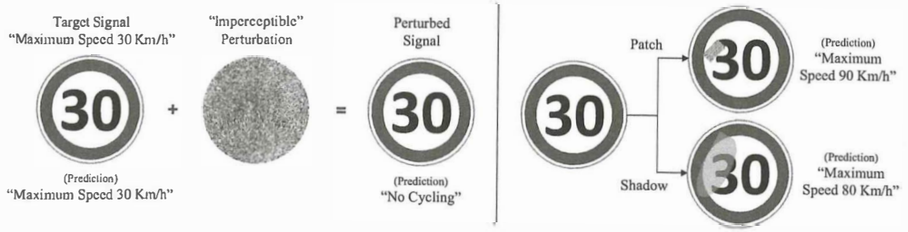
flexible, because it allows providers of AI systems to delineate the required procedures to meet these specifications, the framework imposes a well-detailed ex-ante assessment of the conformity of the product and continuous post-market monitoring and reporting.

IV.B Technical AI Issues: Explainability and Instability

There are two main issues that are currently affecting AI systems and making them poorly predisposed to be assessed through "traditional" standards or procedures because not all their mechanisms are properly evaluable. On the one hand, some AI approaches introduce non-determinism, therefore, they could be unstable, that is, the output of the system may change drastically even with a slight variation of the input (Marrone et al. 2019). On the other hand, the "opacity" of most AI techniques (especially DL ones) makes AI systems extremely challenging to be comprehended and evaluated by human operators. While this can be acceptable in some contexts, the lack of a direct interpretation of choices made can be particularly problematic in safety and/or security-critical applications, such as railways.

Also, instability directly affects the robustness and safety of some AI systems (KR2) as it exposes them to threats such as adversarial attacks (e.g., Ren et al. 2020), that is, approaches aimed at misleading a target CNN by over-imposing an ad-hoc noise on input samples (mostly images). Figure 9.8 shows two representative examples of how a target CNN, in the context of traffic signal recognition, can be "fooled", that is, induced to perform a wrong prediction, by different kinds of perturbations (both malicious and natural) (e.g., Ren et al. 2020; Kumar et al. 2020; Zhong et al. 2022; Yang et al. 2020; Wang et al. 2022). The left part of the picture depicts the effect of an adversarial attack aiming at perturbing the signal to be classified by applying a specific noise pattern that is imperceptible to the human eye but could completely change the output of the system; in this specific example, the assumption is that the CNN has been trained to recognise the "target signal" as "maximum speed 30 km/h" but the perturbation induces the CNN to predict that signal as "no cycling". On the other hand, the right part shows that also malicious (e.g., physical patches) or natural (e.g., shadows) visual perturbations could fool it, inducing it to a wrong classification. Clearly, these are only examples but show the threat of adversarial tacks and the need to cope with AI instability both from technology and standardisation perspectives.

Explainability is also a fundamental aspect to consider as the majority of KRs rely on the possibility of comprehending and analysing AI systems. Over the past few years, many explainable



Source: Authors.

Figure 9.8 Examples of adversarial attacks

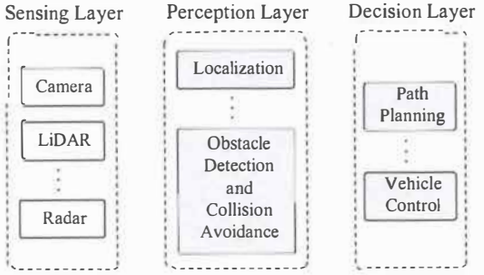
AI (XAI) approaches have been proposed with the aim of unwrapping black-box AI systems and explaining the correlation between the input, the output, and the reasoning of AI models. A sufficient level of explainability, intended as the capability of a model to clarify its functioning (Arrieta et al. 2020), could substantially increase the level of trustworthiness. However, there are some shortcomings that must be considered. By turning a black box into a white box (e.g., a readable model), not only XAI approaches might offer a window towards sensitive data (affecting, de facto, KR3), but they may also create opportunities for adversarial attacks where attackers could benefit from explanations to produce more effective attacks. Furthermore, it has also been shown that adversarial attacks may also strongly affect XAI outcomes (Galli et al. 2021). This latter situation is potentially even more severe, because system supervisors may trust an unreliable system based on the outcomes of a corrupted XAI procedure.

In addition, other complications arise when (i) dealing with multi-modular systems where each module could rely on AI and (ii) when multiple stakeholders are involved (Atakishiyev et al. 2021; Fernández Llorca and Gómez 2021; Omeiza et al. 2021). Notably, by considering an example of a generic and high-level architecture for autonomous vehicles (AVs) similar to that shown in Figure 9.9 (excerpted from Deng et al. 2021), AI could be involved in several subsystems (like those reported within the Perception and Decision layers in Figure 9.9). This means that for a single system based on AI, not only would it require multiple explanations (one for each AI functionality), but each AI functionality is also required to be explained in different manners to allow any stakeholder to comprehend its functioning.

To conclude, several advancements have been made in the recent past towards the *realisation of XAI approaches*. Some frameworks are being developed based on model-specific or model-agnostic techniques. The former deals with the inner working of specific models, while the latter aims to explain the predictions of any models. Some of these approaches are summarised in Linardatos et al. (2020), although well-defined procedures are still required to validate and certify AI-based models.

IV.C Towards Proposals and Standards for AI Systems

In the past few years, several proposals have been presented concerning the assessment and standardisation of AI systems. As an example, the European Union Aviation Safety Agency



Source: Authors.

Figure 9.9 High-level architecture for autonomous vehicles

(EASA) and Deadalean have published a technical report called "Concepts of Design Assurance for Neural Networks (CoDANN) II" (EASA and Daedalean 2021), which defines a W-shaped process to properly assess the usage of neural networks in avionics through a visual traffic detection system as a case study. In addition, EASA also published a concept paper defining some preliminary guidance for the adoption of level 1 AI/ML applications (EASA 2021) – applications oriented towards assisting human operators. These ideas are clearly oriented to avionics, however, they could also be adopted as a valuable starting point in railways.

From the standardisation perspective, it is important to mention that CEN and CENELEC have launched a new technical committee on AI in 2021. The new *CEN/CLC/JTC 21* committee aims to produce, in the near future, standardisation deliverables on AI and related data. It will also take into account the work performed within other international standards and organisations (e.g., ISO/IEC JTC 1/SC 42). Furthermore, the IEEE Standards Association (IEEE SA) has proposed a "Global Initiative on Ethics of Autonomous and Intelligent Systems (AIS)", in the context of which, it issued a landmark paper titled "Ethically Aligned Design" (IEEE 2019) intended to provide recommendations and guidance for standards, certification, and regulation for design and use of AIS. It also includes discussions about the potential harm of AIS to privacy, discrimination, and possible negative long-term effects on societal well-being. Practically, this initiative encompasses a series of standards, which are being developed, assessing transparency of autonomous systems, data-related issues and privacy, and ethically driven robotic, intelligent, and autonomous systems. Regarding the assessment of safety and the trustworthiness of AIS, IEEE is also carrying out activities within "The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)" in relation to bias in AIS. Notably, the bias issue in AI systems is also one of the main topics analysed within the ISO/IEC TR 24027:2021 (ISO/IEC 2021) standard. In addition to the bias consideration, another important related issue, as discussed above, is explainability. In this context, the XAI Working Group of the IEEE Computational Intelligence Society Standards Committee (CIS/SC) is working on the development of XAI standards expected to be submitted to the IEEE SA for Initial Standards Association Ballot in January 2024. Similarly, the ISO/IEC JTC 1/SC 42 technical committee is working on the ISO/IEC AWI TS 6254 standard, which should provide methods and techniques to deal with explainability in ML and AI systems. For a more comprehensive analysis delineating the current panorama on AI-related standards, the reader is referred to Nativi and De Nigris (2021).

V. CURRENT INVESTIGATION TOWARDS SMART RAILWAYS

Figure 9.3 shows classes of problems that researchers have addressed or are currently facing through AI approaches. Even though all of them could bring valuable benefits to the railway system, there are some specific problems that are currently considered hot topics given the huge impact they could have on the rail sector in terms of safety, capacity, punctuality, and energy efficiency. These include the following: smart maintenance of level crossings (LCs); autonomous train driving, specifically virtual coupling (VC) and obstacle detection; and graph embedding-based train delay prediction, as well as incident attribution analysis. Notably, given that AI is a huge domain, by investigating these topics it would also be possible to discern and then outline a set of AI guidelines that could be used to address similar problems in railways. Similarly, it could be possible to analyse what has been developed or

conceived in sectors other than railways (e.g., automotive) and transfer these ideas by properly adapting them to solve railway problems (RAILS 2022a; RAILS 2022b; RAILS 2022c).

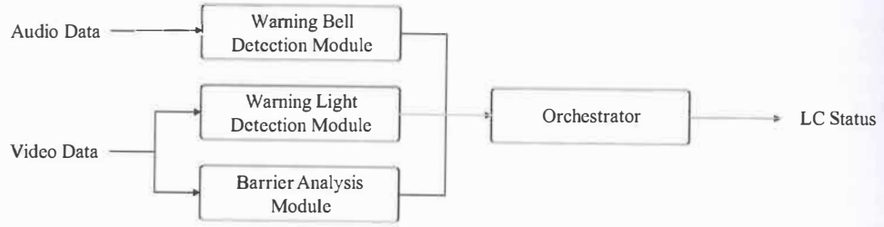
V.A Smart Maintenance at Level Crossings

According to the latest report on "Railway Safety and Interoperability in the European Union" by the European Union Agency for Railways (ERA 2022), in the period from 2016 to 2020, crashes at level crossings accounted for about 30% of all crashes registered in the rail network. Therefore, LCs represent a class of railway assets that raise several concerns in terms of safety, but also maintainability.

Ideally, the most effective way to reduce these crashes would be to substitute LCs with bridges or subways; however, given that it has been estimated that there are more than 100,000 LCs within the EU (ERA 2020), other short-term interventions should be considered to increase safety such as those proposed within the SAFER-LC project (SAFER-LC 2019). These include new warning signalling systems, speed bumps, and also detection and risk evaluation systems based on machine learning. They also include those reviewed by Singh et al. (2021; mostly based on IoT and train-to-wayside communication). Fayyaz et al. (2021) also gave a complete review of suitable obstacle detection technologies and their associated algorithms, which can be exploited to support risk reduction at LCs, and analysed the combination of obstacle detection sensors with intelligent decision layers (e.g., DL models) as an opportunity to provide robust interlocking decisions. These solutions aim to increase the safety level at LCs. However, the other alternative approach is represented by proper maintenance and inspection activities, thus guaranteeing the correct functioning of LCs, which should be the first step in order to ensure safety and capacity availability (RAILS 2022d).

Inspection activities conducted on a scheduled basis may not be so effective as failures may occur between adjacent inspections. Therefore, in the last few years, there has been a growing interest in continuous monitoring applications for LCs to collect data in real-time and promptly detect malfunctions (RAILS 2022d). The main problem, in this case, is related to the installation of IoT sensors representing one of the main enablers for real-time monitoring. Intrusive IoT sensors (i.e., those that are applied directly to the asset to be monitored) may lead to two main complications. First, although the newest LCs may be already equipped with the required sensors, there are multiple LCs deployed in the past decades that may not have them. Hence, a massive, expensive, and time-consuming instrumentation would be required. Second, because LCs are safety-critical assets, the introduction of intrusive external components may lead to an expensive and time-consuming re-approval process. In this context, a suitable solution would be to rely on non-intrusive and cost-effective sensors such as cameras and microphones. Notably, cameras may be already installed at some LCs for surveillance purposes. In those cases, there will not be any additional costs besides the development of AI models capable of retrieving the health status of LC components from video and audio data.

Leveraging the research activities conducted within RAILS (2022d), Figure 9.10 shows a general AI framework for the non-intrusive analysis of the health status of LCs. Each module oversees a particular LC subsystem (i.e., barriers, warning bells, and warning lights). Therefore, the warning bell detection (WBD) module detects and analyses the functioning of the warning bells. Similarly, the warning light detection (WLD) module monitors warning lights. Then, the barrier analysis (BA) module analyses the behaviour of the barrier. Lastly, the orchestrator gathers the output of the previous modules and outlines the health status of



Source: Authors.

Figure 9.10 *AI framework for non-intrusive level crossing monitoring*

the LC. Notably, the implementation of the different modules in terms of AI approaches is independent of the specific LC, which means that they could be adapted to any kind of LC.

Possible AI techniques that could be leveraged to build such modules are discussed next.

- The detection of warning bells could be seen as a classification task involving three classes: warning bell, background noise (which practically involves all the sounds that are not alarms/sirens), and generic alarm (which involves all the possible alarms and sirens that could be recorded near level crossings and can be confused with warning bells). Then, given that it would be possible to pre-process the audio to retrieve images related to their spectrum, it would be possible to adopt CNNs to classify them. As a possible solution, the VGGish architecture (Hershey et al. 2017), originally developed to deal with AudioSet data (Gemmeke et al. 2017), can be adapted to this task. Notably, AudioSet is a dataset composed of multiple records of various events and could be used to pre-train the chosen AI model (in the context of a transfer learning approach) or even retrieve some samples for the "background noise" and "generic alarm" classes. Then, based on the predictions of the CNN, it would be possible to understand whether the warning bell is working correctly or not.
- As for the WLD module, the peculiarity of this system is that, given that the camera is fixed, the warning lights will always occupy the same portion of the images unless external agents (like the weather) move the camera. Therefore, the easiest approach would be to pre-process the images and crop out the warning lights. A classifier can then be implemented to classify the lights. By leveraging the predictions, the AI model would understand if the lights are blinking correctly or not.
- Lastly, the BA module should understand if the barrier is moving properly (e.g., smoothly) or not. To do that, it would be possible to leverage some DL-based object detectors (e.g., YOLOv4 (Bochkovskiy et al. 2020)) to detect the barrier in the images over time. These models would produce a bounding box, that is, a square that almost perfectly contains the barrier. Hence, it would be possible to plot, for example, the heights of subsequent bounding boxes to retrieve the "path" that the edge of the barrier traces over time. Then, it would be possible to analyse the path by comparing it with a reference path to retrieve a score for the health status of the barrier. The best approach to compare these paths should be defined through a results-guided analysis because these paths can be conceptually seen as signals and, most likely, the majority of DL and traditional ML techniques could produce suitable results.

V.B Autonomous Train Driving

Before analysing possible solutions for autonomous train driving (ATD), it is necessary to define what autonomous means, and what is the difference between automation and autonomy. *Automation* refers to the capability of a system to automatically or semi-automatically perform specific tasks based on pre-specified rules (Flammini et al. 2022). Differently, *autonomy* refers to the ability of a system (a train, in this case) to dynamically adapt to unexpected scenarios by taking independent decisions (Milburn and Erskine 2019) where the main enabler is AI (Fernández Llorca and Gómez 2021). According to these definitions, and the one given by Beš inović et al. (2022) for AI (which was reported in Section II), current driverless trains, which rely on systems such as ATO and ATP, cannot be classified as autonomous or intelligent since they lack any learning and adaptation capabilities (Flammini et al. 2022).

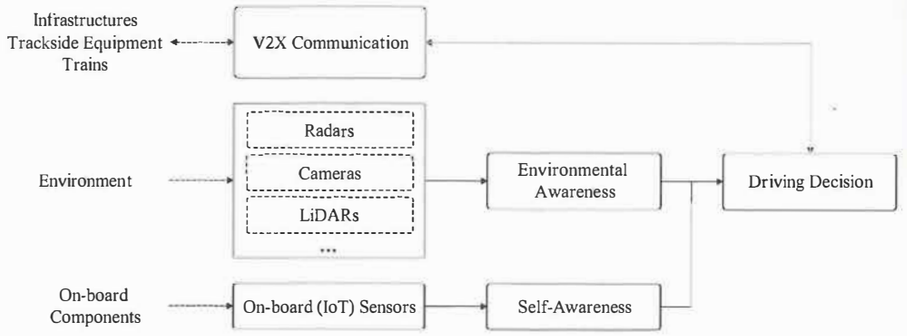
The importance of introducing AI in train driving comes from the fact that despite current automatic trains showing a great impact in segregated environments (in terms of system capacity, punctuality, and energy consumption), to achieve the same benefits in open railway environments, automation may not be sufficient. Although ATP systems are affirmed in both segregated and open environments, ATO implementation is still experimental in open railways given their mutability. It is worth highlighting that some European projects conceived in the context of the S2R Innovation Program 2 (IP2, "Advanced Traffic Management and Control Systems") are actively working on the integration of ATO within the European Train Control System (ETCS) to achieve the so-called ATO over ETCS (AoE). Valuable results have been obtained for AoE up to GoA2 (Grade of Automation 2) (X2RAIL-1 2016) and ongoing research is exploring solutions towards GoA3/4 AoE (X2RAIL-4 2019). However, AI, if properly implemented, could be extremely valuable in these contexts because it could efficiently increase trains' awareness.

From a conceptual perspective, autonomous trains (AT) should be capable of taking autonomous decisions based on the health status of their components (self-awareness), the status of the environment in the surroundings (environmental awareness), and communications coming from other ATs or infrastructures (V2X communication). Figure 9.11 graphically connects these aspects.

In addition to solutions oriented at evaluating the health status of train components – which can be considered as maintenance and inspection approaches with some more trustworthy concerns related to their application to safety-critical functionalities – they are next discussed as AI applications that could bring some benefits in exploiting V2X communications and analysing the environment and are also given some consideration regarding the trustworthiness of autonomous vehicles.

V.B.1 Cooperative driving for virtual coupling

In the field of ATs, the S2R programs (MOVINGRAIL 2020) and (X2RAIL-3 2018) proposed a novel paradigm based on the concept of virtual coupling (VC). The idea, strictly related to the platooning concept in the automotive field, is to virtually couple two or more trains via train-to-train (T2T) communication so they can travel in formation with the same velocity while maintaining a desired inter-train safety distance, in order to allow them to run at a closer distance than the absolute braking distance of the rear consist, that is, the group of rail vehicles that make up a virtually coupled train set (VCTS).



Source: Authors.

Figure 9.11 AI framework for autonomous trains

The VC paradigm could bring a broader set of advantages over the traditional way to operate a railway network, such as improving infrastructure utilisation, increasing the capacity of the existing railway lines (reduction of trip times, headways, etc.) with respect to current systems, increasing flexibility and allowing platooning among trains of different types, reducing costs, increasing competitiveness, and making more efficient goods and passengers transportation with respect to road transportation.

Few contributions regarding AI approaches for vehicle platooning have been proposed so far that show promising results when ML methods are exploited (RAILS 2022a). In particular, it has emerged that reinforcement learning (RL) methods could outperform, especially in complex and uncertain scenarios, conventional model-based approaches, and thus could be considered for implementation in railway VC. Indeed, an RL approach could allow the consist comprising the platoon to learn how to interact with unknown environments (without any prior knowledge of the latter) via a learn-by-doing process where it could learn, through trial and error, the best way to accomplish a task. The surrounding environment would embed all the unpredictable and uncertain factors characterising railway dynamic and complex scenarios, such as track conditions (e.g., adhesion factors, gradients), exogenous factors, heterogeneous trains with different operational performances (e.g., different braking capabilities, different speed categories), trains with variable mass (e.g., freight trains), and uncertainties in train location information. Note that the consists themselves and their operational performances would be embedded in the environment. This means that RL approaches could allow the coupling of trains with different operational capabilities, ensuring platooning among heterogeneous trains while absorbing uncertainties arising in real-world driving conditions.

In order to ensure the effectiveness of VCTS, the technical performance requirements for T2T communications (direct T2T over short and long distances, low latency, etc.) should be satisfied. As emerged from MOVINGRAIL (2020), 5G technology could meet the T2T communication requirements for VCTS.

One of the main issues related to the effectiveness of RL approaches is the learning process. Namely, RL methods work in high-dimensional, continuous action spaces to deal with physical control tasks. Such large action spaces are difficult to explore efficiently, and no dataset would be adequate for the purpose. That is why simulators for virtual validation and training

are required. In particular, in order to be exploited for railway VC, they should take into account T2T communications, as also pointed out in RAILS (2022e).

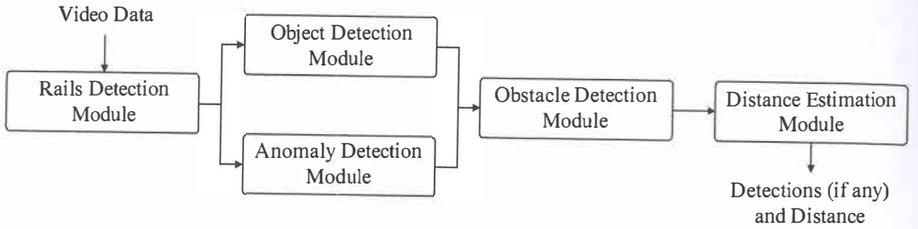
V.B.2 Obstacle detection and collision avoidance

In the context of ATs, obstacle detection is one of the possible applications that could play a central role in improving safety as it could give trains the ability to "see" what is on the rail tracks and understand if they can safely proceed or should take some preventive actions to mitigate or possibly avoid possible collisions.

The problems that characterise the rail sector with respect to others, for example, automotive, are mainly related to: (i) the speed of the vehicles; (ii) the inflexibility of the path, indeed, there is not any possibility to avoid obstacles if not by braking; and (iii) the adhesion between wheels and rails, which is less than that between tyre and asphalt (RAILS 2022e). On the other hand, there are some advantages related to the mutability of the environment. Indeed, rails can be considered more "closed" than roads, which means that, by excluding level crossings and stations, it is not so common to find obstacles on rail tracks. Also, the trains always follow the same trajectory, which gives the rail track a sort of regularity over time. Nevertheless, complex and multi-modular systems are required to detect obstacles at acceptable distances to mitigate potential crashes. It is worth noting here that the S2R projects (SMART 2016; SMART2 2019) have worked on a complex multi-modular system for obstacle detection involving three main subsystems – onboard, airborne, and trackside – and different sensors like RGB cameras and lasers.

Usually, when it comes to detection systems, vision-based approaches are mainly oriented at classifying potential obstacles – that is, they associate each detected obstacle with a label (e.g., car, human) rather than identifying potential anomalies in general (typically demanded in sensors like radars or LiDARs). To be more specific, most vision-based DL systems are trained in a supervised manner, which means that they are trained on a set of labelled data, and they would most likely not be capable of detecting obstacles that have not been included in the labelled set. This is an extremely sensitive issue as it would be nearly impossible to take into account a priori every object that could occupy rail tracks. Therefore, for the sake of robustness and, to some extent, trustworthiness, it would be advisable to extend the effectiveness of vision-based DL systems by implementing a multi-modular architecture as depicted in Figure 9.12 (RAILS 2022e).

Notably, the framework adds an anomaly detection (AD) module to those typically considered when addressing vision-based obstacle detection, which includes the rails detection (RD) and the object detection (OD) modules (Ristrić-Durrant et al. 2021a). Briefly, the RD module identifies the tracks and then the OD and the AD modules detect the objects (i.e., obstacles known a priori) and potential anomalies (i.e., any kind of obstacles including those unknown a priori) on rail tracks, respectively. Hence, the check detection (CD) module merges the predictions while the distance estimation (DE) module estimates the distance of the object and/or the anomaly. The main difference between the OD and the AD modules is that the first one is trained in supervised mode, as explained above, while the second one is trained in unsupervised mode. Visually, the AD module would be capable of learning the characteristics of "free" tracks (rail tracks without obstacles) to highlight at run-time all the possible anomalies, that is, regions of pixels within the images that have different characteristics with respect to free tracks. The advantages of introducing the AD module, if all the AI approaches are properly implemented, include:



Source: Authors.

Figure 9.12 *A general framework for vision-based obstacle detection*

- The AD module would act as a complementary system for the OD module. The prediction performed by the AD could be used by the CD to confirm the detection coming from the OD.
- The AD module could also act as a fall-back system and should be capable of detecting anomalies even though the OD module does not detect any object.
- The AD module would increase coverage of the vision-based subsystem as this would be capable of detecting any kind of obstacle and classifying those known a priori.
- The whole vision-based subsystem, up to a given distance, can be used in a modular redundancy architecture to check anomalies detected by means of other sensors like radars or LiDARs.

As for the possible techniques that could be involved to implement these modules, Ristić-Durrant et al. (2021) carried out a comprehensive review of traditional and AI-based CV approaches that could be used to implement the RD and OD modules. From a DL perspective, semantic segmentation approaches, for example, U-NET (Ronneberger et al. 2015) could be extremely useful to detect rails, while object detectors like those belonging to the YOLO and region-based CNN families (Zou 2019) have shown high performances in detecting objects even in real-time. As for the AD module, it seems that DL techniques have not yet been fully addressed to deal with this specific problem in railways. However, deep autoencoders seem to be a suitable solution to this problem because it would be possible to train them with defect-free images (i.e., free tracks) and then obtain an anomaly map in the output to identify the potential anomaly (e.g., Bergmann et al. 2018). Lastly, although applications related to the AD and DE modules have not been fully investigated yet, some useful hints and primary solutions suitable for estimating the distance of an object have been proposed by Ristić-Durrant et al. (2021b). The estimation of the distance of anomalies and the size of an anomaly remain open issues that are ripe for future research.

V.B.3 Trustworthy autonomous trains

As mentioned in Section IV., when AI is directly applied to control safety-critical systems (e.g., a train's driving system), major trustworthiness and regulatory concerns arise. Section IV also presented an overview of the European vision of trustworthy AI and introduced AI-HLEG's ALTAI framework. In this context, it is worth mentioning that Borg et al. (2021) analysed the

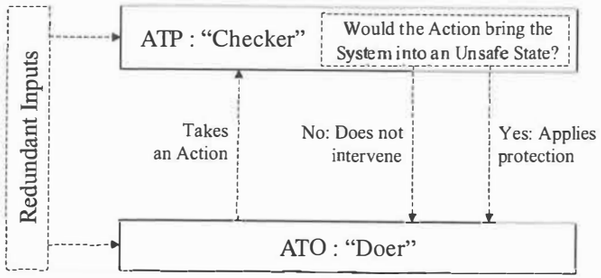
applicability of the ALTAI framework to advanced driving-assistance systems (ADAS) based on ML. The authors found that the ALTAI is largely applicable to ADAS, however, there might be room for some improvements in order to make ALTAI even more suitable to these systems (including ALTAI variants for each development phase) and different contexts. Along this line, the "Trustworthy Autonomous Vehicles" (Fernández Llorca and Gómez 2021) report aimed to map the ALTAI requirements (and criteria) to autonomous vehicles. Similarly, with in-depth analyses, it would be possible to precisely map the ALTAI requirements to autonomous trains to have some general guidelines towards the design and implementation of AI systems for trustworthy autonomous trains.

In general, the main concerns derive from the current level of the maturity of AI systems and the relative absence of specific regulations and standards that could help to identify the safety integrity level (SIL) of AI systems. This is probably directly related to the explainability and instability issues of AI approaches as discussed in Section IV.B. Indeed, it is extremely challenging to understand "why" and "how" some AI models (especially DL ones) make a specific decision. Consequently, it is difficult to comprehend the cause-effect relationship between a given input and the related output, also because the instability may make this correlation even more challenging to detect. These problems, among others, make AI systems extremely complex to certify or even analyse. It is however worth noting that, as it emerges from the latest reviews on XAI for autonomous driving in the automotive industry (Atakishiyev et al. 2021; Fernández Llorca and Gómez 2021; Omeiza et al. 2021), most of the post-hoc XAI approaches (i.e., explanations models applied a-posteriori on the AI system) may have an impact on the safety and trustworthiness of autonomous vehicles. Furthermore, concrete regulations on autonomous vehicle explainability are still missing, and limited work has been done in the context of explainable localisation. Overall, the application of these methods in the field of AVs is still at a nascent stage and needs to be more deeply investigated (the same applies to explainable control).

Having stated that, it is trivial to deduce that AI applications like obstacle detection are challenging to be physically implemented given that, at the moment, it seems not to be possible to establish their SIL. However, some interesting suggestions have been proposed by Flammini et al. (2022), where the concept of the safety envelope (Koopman and Wagner 2018) has been highlighted and could be adopted to allow AI systems to control trains' movements within the safety envelope area that is free from any risk of collisions and other hazards, which is continuously computed and updated. In particular, it would be possible to exploit the ATP system (which can be currently certified as SIL4 in both segregated and open environments) as the safety envelope.

Figure 9.13 shows the relationship that currently holds between ATP and ATO based on the concept of the safety envelope. Practically, the ATP is responsible for the safety and correct application of the dynamic speed profile to avoid consequences such as derailments and collisions, and supervises the actions taken by the ATO. Hence, in case the ATO, for any reason, performs an action that would bring the system into an unsafe state, the ATP will apply the adequate countermeasures to protect the system (e.g., bring it into a fail-safe state).

In the same way, as long as the ATP is available, it could supervise the decisions taken by an intelligent train operation (ITO), which is expected to extend ATO functionalities by introducing an adaptive behaviour based on AI to optimise energy, capacity, and comfort. Notably, this would increment the trustworthiness of systems including intelligent functionalities because these are controlled by non-intelligent systems in a sort of system-over-the-loop paradigm.



Source: Authors.

Figure 9.13 *ATP as ATO's safety envelope*

Additionally, certification against reference safety standards using existing or traditional approaches would also be possible at the system level because safety concerns demanded by the ATP do not involve any intelligent behaviour.

To conclude, deep RL approaches seem to be optimal candidates to support the development of ITO functionalities. Those models have already been investigated in the automotive industry (Kiran et al. 2021) and could also be exploited in railways to ensure that trains adapt to constraints (like braking curves) while providing optimal driving policy (e.g., Huang et al. 2019).

V.C Smart Railway Traffic Planning and Management

V.C.1 Graph embedding-based train delay prediction

Dissatisfaction among passengers and significant financial loss might result from train delays. The framework of the train delay prediction problem has been used to study a huge range of solutions. Effective prediction depends on understanding how to appropriately describe specific train characteristics. One of these characteristics that is difficult to properly interpret, for instance, is the path of a train, which includes the origin, intermediate stations, and destination. This is because the route of a train is topologically complex and too abstract to be represented by traditional data structures. In this section, a case study is introduced using graph embedding approaches to comprehend and simulate the intricate railway system structure from a network perspective. The case study aims to capture a wide range of elements, such as network topology, infrastructure information, and train profile. To embed a train's path in a network topology perspective, a method based on structural deep network embedding (SDNE) and singular value decomposition (SVD) is presented for the first time. The resulting route embedding is very accurate and dependable in terms of capturing the network architecture and requires substantially less computational effort than principal component analysis (PCA), which is another conventional way to encode and condense the information for complex characteristics. In contrast to computational models based on PCA, the graph embedding-based models are competitive in terms of prediction accuracy and are significantly efficient in computational time.

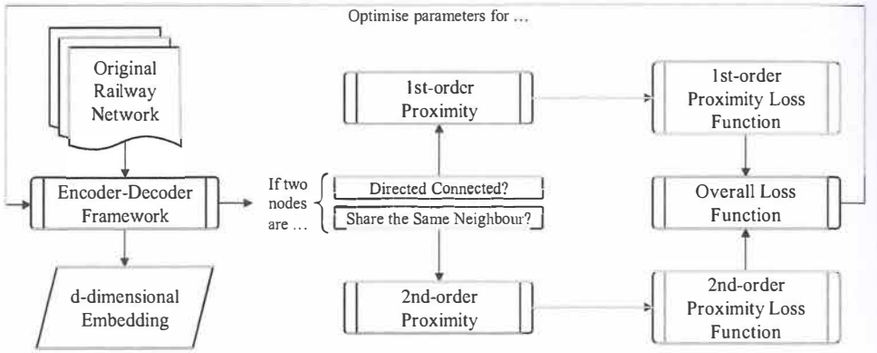
The objective of the delay prediction case study is to forecast the total/summed delay level for an unobserved train service based on delay data collected from railway operations history,

taking into account the characteristics of individual railway stations as well as the structural network characteristics (e.g., connectivity between these stations, general travelling time for a given link, network density for different areas). It is crucial to understand the network links between stations since they can influence the delay status for train services at a certain railway station. That is, it is shown that the closer a railway station is to its neighbouring stations geographically, the more likely that a delay will affect those stations.

As a successful dimensionality reduction method in the field of computer science, Wang et al. (2016) developed the SDNE graph embedding algorithm. In order to preserve the structural relationships in the original network, the fundamental concept behind embedding is to keep related nodes close to one another in the vector space. Although each value in the vector cannot be explained, it nevertheless partially reflects the traits of a specific station. Such a representation is helpful for comparing the similarity between two stations. Given that the SDNE approach considerably reduces the amount of essential information, vector operations can be carried out more quickly and easily than using traditional mathematical procedures. However, the literature available today suggests graph embedding is an emerging technology that has only been applied on assisting bundle mining for online shopping and link prediction tasks. There is no evidence showing that this technology has been successfully adapted to the public transport industry, let alone rail traffic planning and management. It is a promising direction to combine SDNE and SVD together to predict train primary delays. The major contributions can be summarised as follows:

- To conduct train delay prediction tasks within the context of the UK national rail network, both spatial network characteristics and historical traffic dynamics are integrated into one framework.
- A deep neural network-based graph embedding technique is also used to extract network properties both globally and locally. Conventional machine learning algorithms have been deployed as solutions for train delay prediction.
- According to the authors, this is the first attempt to process complicated features for network objects by integrating matrix decomposition technology with graph embedding methodologies. It also successfully integrated route information for train services from a network topology perspective.

The SDNE module in the overall methodology framework is to acquire structural deep network representation. The original SDNE method has been refined and localised to better fit the context of the railway network structure such that stations in the railway network can be effectively represented in Figure 9.14. Only node embedding vectors can be produced for each station under the SDNE model. Route embedding, which may be used to describe each train's unique route feature and is crucial for characterising railway services, is another significant aspect that needs to be given more attention. Since route lengths might vary greatly, it is not applicable to simply focus on the node vectors of the stations that a train passes through. Additionally, doing so would result in extremely large vectors that would be unusable for representing routes, much like one-hot encoding. For the purpose of producing route vectors, singular value decomposition (SVD) is introduced. SVD is an effective tool for retrieving matrix information (Yang et al., 2011). By extracting the most important data from the original vector (referred to as "singular values"), the original data container by a much compressed data structure using SVD is generated, which dramatically reduces noise and redundant data. The



Source: Authors.

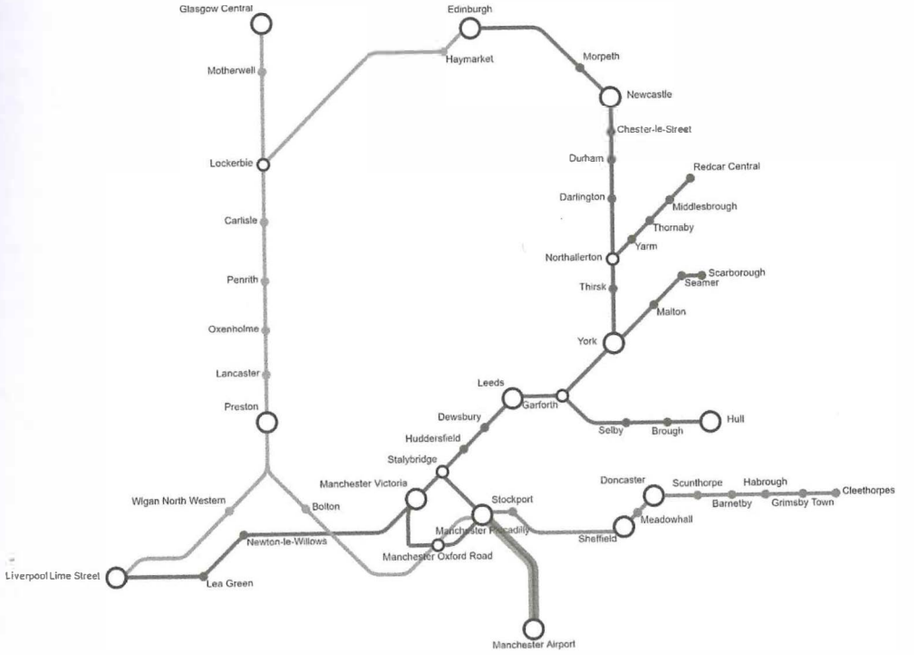
Figure 9.14 High-level architecture of the refined SDNE model

decomposition process generates a matrix Σ with only diagonal elements. One of the important characteristics of this matrix is that its diagonal elements are arranged from the largest to the smallest value. This case study (applied on the network shown in Figure 9.15) uses the singular value matrix to represent the route embedding vectors, allowing the route lengths of various train services to be normalised, and the dimensionality of the original network features to be significantly reduced without losing any crucial topology structure information.

To demonstrate the feasibility of the approach, the graph embedding described above is applied to the TransPennine Express (TPE) rail network. TransPennine Express, the data source provider, is a well-known British train operating company that runs rail services between the cities of Northern England and Scotland. The three primary routes connect major cities including Glasgow, Liverpool, Leeds, and Newcastle, and these services cover three regional routes in the vicinity of Manchester (shown in Figure 9.15). TPE provided information on train services and delays for the time periods of 28 May 2017 through 24 June 2017, and 27 May 2018 through 23 June 2018. The network consists of 177 stations, 192 edges/links between the stations, and 1,191 train delay instances running within the network. The dataset that was obtained includes network features, label features, categorical characteristics, numerical features, and temporal features. In order to predict primary train delays, three traditional supervised ML algorithms were fed the output of the route-embedding implementation, which was completed after processing other influencing parameters, such as departure/arrival time, passenger volume, total margin time, speed limit, and rolling stock type. Decision tree (DT) (Quinlan, 1987), random forest (RF) (Ho, 1995), and multi-layer perceptron (MLP) were the three ML algorithms used to test the SDNE+SVD architecture (Gardner and Dorling, 1998).

V.C.2 ML-based incidents attribution analysis

A number of previous studies assessed the utilisation of intelligent strategies and AI-based techniques to derive the occurrence of railroad incidents/events. Making such forecasts begins with gathering useful information from insightful incident reports. Every event occurs either in a station or on a segment of rail track, and each of these holds a set of contributing



Source: Authors.

Figure 9.15 Considered TPE rail network

factors. These causes, whether they solely or collectively cause the incident, must be well-documented as soon as possible once the incident is detected. The specific internal/external factors that determine the occurrence of disruptions can be measured or formulated using a variety of methodologies. The maximum likelihood estimation is most widely used since it is founded on actual facts rather than technical judgements (Chen et al., 2022) by revealing the frequency, distribution, and co-occurrence of each crash text phrase. For instance, Syeda et al. (2019) proposed an NLP model to extract insights from railway crash data. Their study demonstrated how text analysis technologies may be used to map various English text sequences to concepts and entities in incident domains. In another study, word embedding techniques were used (Heidarysafa et al., 2018) regarding the relationship between crash report texts and the actual causes of adverse occurrences (GloVe and Word2Vec). The findings demonstrated that this method could identify crash causes with accuracy based on narratives in report files and can also identify significant discrepancies between recorded reporting and actual conditions.

Incorporating empirical data into the statistical analysis/regression process inherently has its advantages in the tasks of abnormal events/disruption prediction. However, considering empirical data to evaluate disruption frequencies and implications of various disruptions is insufficient. The development of a supervised learning prediction model to forecast disruption

frequency and impact for each distinct element of the rail transport network is both necessary and promising because it enables disruption predictions to be made at any given station for a given time period without the need for an adequate number of empirical disruption observations at each location and time interval. In addition, using ML-based techniques can make the disruption impact prediction process more effective, particularly when there are many different disruption cases. As a result, the supervised learning approach has the advantage of resolving the computational issues caused by prevalent traffic simulation models used in actual railway networks.

With the context of ensuring the safety of rail traffic, some effort has been done to investigate incident frequencies. The fact that most of these traffic studies largely used descriptive and aggregate models to forecast probabilities of different incidents is one of the objectives' shortcomings. On the one hand, because of the naturally complicated nonlinear spatial-temporal links or interactions between traffic participants, the pattern of disruption propagation and the occurrence of primary delays are both largely determined by these factors. If these relationships are not maintained, it will be difficult to predict them accurately in practical situations. However, doing so would result in a computing space that is excessively large because all observed relations and deterministic elements, including graphical objects, would be fed straight into a traditional statistical analysis function or descriptive model. To do this, a disaggregated modelling approach to forecast interruption frequencies is proposed to forecast the accompanying impacts for each disruption type. To make these predictions, a supervised learning strategy is proposed, which enables the predictor to calculate disruptions at specific stations for each time period. This approach allows for a quick and accurate prediction of disruption impacts for more unknown disruption instances without necessitating a large number of historical disruption observations because there are insufficient empirical disruption observations available for each location and time slot. Scientifically, by taking into account the unique characteristics of each station, well-known supervised machine learning models like logistic regression, MLP, KNN, and random forest are qualified to predict the frequency of occurrence for various types of accidents and their impact on passenger delay for individual railway stations. In order to support the dispatch agent in prioritizing areas where mitigation measures (rescheduling) are needed, it is important to present the agent with expected disruption impacts for each unique station in the network and each distinguishing time period and disruption type. The location, timeframe, and date of crashes and disruptions are the module's intended outputs.

Network Rail would regard primary delay as a scalar that mainly contributes to the performance of train services, even though there are various possible reasons for a particular railway incident. For example, in the section of "External events", a category of different types of incidents causes are summarised in Table 9.5.

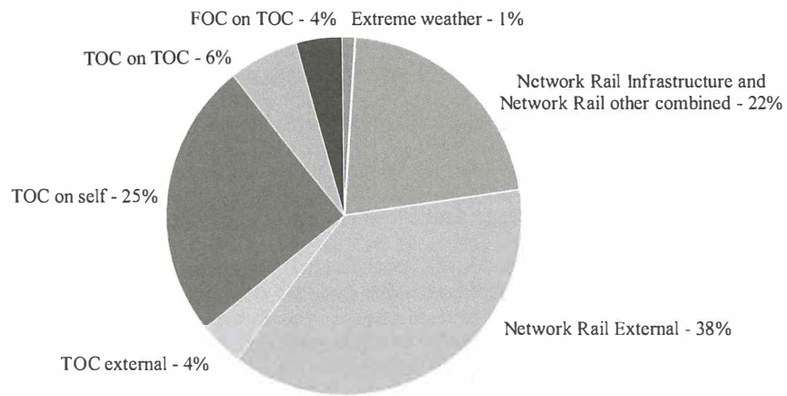
And according to Network Rail 2022, disruption responsibility corresponding to the British railway transport system can be attributed to different causes (see Figure 9.16).

In the diagram, particulars are shown including operational issues and damage to or failure of the infrastructure of the railway such as tracks, signalling, or points, including where bad, but not extreme, weather causes delays to the rail network (Network Rail responsibility).

- Network Rail external are the delays caused by external factors such as weather, trespass, vandalism, cable theft, and fatalities.
- TOC external includes issues that the train company could not have foreseen such as passenger action or illness on trains, damage to trains by road vehicles, or the forced closure of a station

Table 9.5 *Types of incidents causing delays*

Delay attribution codes	Explanation of incident types
A*	Freight terminal operation causes
D*	Holding codes
F*	Freight operation causes
I* and J*	Infrastructure causes
M* and N*	Mechanical or fleet engineer causes
O*	Network rail operating causes
P*	Planned or excluded delays or cancellations
Q*	Network rail non-operating causes
R*	Station operating causes
T*	Passenger operating causes
V*	External events – TOC responsibility
X*	External events – network rail
Y*	Reactionary delays
Z*	Unexplained delays and cancellations



Source: Authors.

Figure 9.16 *Percentages for different disruption responsibilities*

- TOC on TOC: one train operating company having their services delayed by the actions of another train company such as a delayed train from one company causing other trains to be delayed (delay propagation).
- FOC on TOC: train companies having their services delayed because of the actions of a freight train on the network. (Freight trains are not included in this case.)

VI. CONCLUSIONS

The railway system is a transportation sector where the increase in performance and safety in the near future will depend on the ability to derive insights from complex datasets and use that information for monitoring, controlling, and taking optimal decisions both beforehand and in real time. AI technologies have the potential to streamline operational performance by increasing the level of effectiveness in decision making and improving overall efficiency. For rail, this is an opportunity deserving of further expansion, while coping with several challenges. This chapter provides a comprehensive overview of the current state of play, spanning from the opportunities and the current applications of AI to railway problems, to the analysis of the main issues to overcome, and the regulatory horizon with a focus on European standards, laws, and guidelines. Then insights were given into some ongoing research activities to highlight how AI could bring benefits in addressing railway problems related to maintenance, intelligent control, and traffic planning and management, also including visionary applications such as cooperative driving and virtual coupling.

REFERENCES

[1] AI HLEG (2019). Ethics Guidelines for Trustworthy AI. European Commission, B-1049 Brussels. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 4 October 2022.

[2] AI HLEG (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. European Commission, B-1049 Brussels. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment. Accessed 4 October 2022.

[3] Antoniou, A., Storkey, A., & Edwards, H. (2017). Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340.

[4] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.

[5] Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2021). Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. arXiv preprint arXiv:2112.11561.

[6] Banerjee, M., & Pal, N.R. (2014). Feature selection with SVD entropy: Some modification and extension. *Information Sciences*, 264, 118–134.

[7] Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., & Steger, C. (2018). Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011.

[8] Bešinović, N., De Donato, L., Flammini, F., Goverde, R.M.P., Lin, Z., Liu, R., Marrone, S., Nardone, R., Tang, T., & Vittorini, V. (2022). Artificial intelligence in railway transport: Taxonomy, regulations, and applications. *IEEE Transactions on Intelligent Transportation Systems*, 23(9), 14011–14024. doi:10.1109/TITS.2021.3131637.

[9] Bochkovskiy, A., Wang, C.Y., & Liao, H.Y.M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[10] Borg, M., Bronson, J., Christensson, L., Olsson, F., Lennartsson, O., Sonnsjö, E. Ebabi, H., & Karsberg, M. (2021). Exploring the assessment list for trustworthy ai in the context of advanced driver-assistance systems. In 2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics) (pp. 5–12). doi: 10.1109/SEthics52569.2021.00009.

[11] Chenariyan Nakhaee, M., Hiemstra, D., Stoelinga, M., & van Noort, M. (2019). The recent applications of machine learning in rail track maintenance: A survey. In Collart-Dutilleul, S., Lecomte, T., Romanovsky, A. (Eds.), *Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification*. RSSRail 2019. Lecture Notes in Computer Science, 11495. Cham: Springer. https://doi.org/10.1007/978-3-030-18744-6_6.

[12] Chen, Z., Huang, K., Wu, L., Zhong, Z., & Jiao, Z. (2022). Relational graph convolutional network for text-mining-based accident causal classification. *Applied Sciences*, 12(5), 2482.

[13] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213–3223).

[14] De Donato, L., Flammini, F., Marrone, S., Mazzariello, C., Nardone, R., Sansone, C., & Vittorini, V. (2022). A survey on audio-video based defect detection through deep learning in railway maintenance. *IEEE Access*, 10, 65376–65400. doi: 10.1109/ACCESS.2022.3183102.

[15] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). doi: 10.1109/CVPR.2009.5206848.

[16] Deng, Y., Zhang, T., Lou, G., Zheng, X., Jin, J., & Han, Q.-L. (2021). Deep learning-based autonomous driving systems: A survey of attacks and defenses. *IEEE Transactions on Industrial Informatics*, 17(12), 7897–7912. doi:10.1109/TII.2021.3071405.

[17] Dirnfeld, R., De Donato, L., Flammini, F., Azari, M.S., & Vittorini, V. (2022). Railway digital twins and artificial intelligence: Challenges and design guidelines. In Dependable Computing – EDCC 2022 Workshops. EDCC 2022. Communications in Computer and Information Science, 1656. Cham: Springer. doi: 10.1007/978-3-031-16245-9_8.

[18] EASA (2021). First usable guidance for Level 1 Machine Learning applications. EASA Concept Paper. https://www.easa.europa.eu/en/easa-concept-paper-first-usable-guidance-level-1-machine-learning-applications-proposed-issue-01pdf. Accessed 4 October 2022.

[19] EASA & Daedalean (2021). Concepts of Design Assurance for Neural Networks (CoDANN) II. https://www.easa.europa.eu/en/document-library/general-publications/concepts-design-assurance-neural-networks-codann-ii. Accessed 4 October 2022.

[20] ERA (2020). Report on Railway Safety and Interoperability in the EU 2020. https://www.era.europa.eu/library/corporate-publications/safety-and-interoperability-progress-reports_en. Accessed 5 October 2022.

[21] ERA (2022). Report on Railway Safety and Interoperability in the EU 2022. https://www.era.europa.eu/library/corporate-publications/safety-and-interoperability-progress-reports_en. Accessed 5 October 2022.

[22] European Commission (2016). General Data Protection Regulation (GDPR). https://gdpr.eu/tag/gdpr/. Accessed 4 October 2022.

[23] European Commission (2018). Communication from the Commission: Artificial Intelligence for Europe. Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.

[24] European Commission (2021). Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Accessed 4 October 2022.

[25] Fayyaz, M.A.B., Alexoulis-Chrysovergis, A.C., Southgate, M.J., & Johnson, C. (2021). A review of the technological developments for interlocking at level crossing. Proceedings of the Institution of Mechanical Engineers, Part F: *Journal of Rail and Rapid Transit*, 235(4), 529–539. doi: 10.1177/0954409720941726.

[26] Fernández Llorca, D., & Gómez, E. (2021). Trustworthy Autonomous Vehicles. EUR 30942 EN, Publications Office of the European Union, Luxembourg. doi: 10.2760/120385.

[27] Flammini, F., De Donato, L., Fantechi, A., & Vittorini, V. (2022). A vision of intelligent train control. In Collart-Dutilleul, S., Haxthausen, A.E., Lecomte, T. (Eds.), *Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification*. RSSRail 2022. Lecture Notes in Computer Science, 13294. Cham: Springer. doi: 10.1007/978-3-031-05814-1_14.

[28] Galli, A., Marrone, S., Moscato, V., & Sansone, C. (2021). Reliability of eXplainable artificial intelligence in adversarial perturbation scenarios. In Pattern Recognition. ICPR International Workshops and Challenges. Lecture Notes in Computer Science, 12663. Cham: Springer. doi: 10.1007/978-3-030-68796-0_18.

[29] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14–15) 2627–2636.

[30] Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 776–780). doi: 10.1109/ICASSP.2017.7952261.

[31] Ghofrani, F., He, Q., Goverde, R.M.P., & Liu, X. (2018). Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*, 90, 226–246. doi: 10.1016/j.trc.2018.03.010.

[32] Heidarysafa, M., Kowsari, K., Barnes, L., & Brown, D. (2018). Analysis of railway accidents' narratives using deep learning. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1446–1453). IEEE.

[33] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 131–135). doi: 10.1109/ICASSP.2017.7952132.

[34] Ho, T.K. (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278–282). IEEE.

[35] Huang, J., Zhang, E., Zhang, J., Huang, S., & Zhong, Z. (2019). Deep reinforcement learning based train driving optimization. In 2019 Chinese Automation Congress (CAC) (pp. 2375–2381). doi: 10.1109/CAC48633.2019.8996988.

[36] IEEE 2019. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. https://standards.ieee.org/industry-connections/ec/eadle-infographic/. Accessed 4 October 2022.

[37] ISO/IEC 2021. Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making. https://www.iso.org/standard/77607.html. Accessed 4 October 2022.

[38] Kiran, B.R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A.A., Yogamani, S., & Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6), 4909–4926. doi: 10.1109/TITS.2021.3054625.

[39] Koopman, P., & Wagner, M. (2018). Toward a framework for highly automated vehicle safety validation. SAE Technical Paper. doi: 10.4271/2018-01-1071.

[40] Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital Twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, 51(11), 1016–1022. doi: 10.1016/j.ifacol.2018.08.474.

[41] Kulkarni, R., Dhavalikar, S., & Bangar, S. (2020). Traffic light detection and recognition for self driving cars using deep learning. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1–4). doi: 10.1109/ICCUBEA.2018.8697819.

[42] Kumar, K.N., Vishnu, C., Mitra, R., & Mohan, C.K. (2020). Black-box adversarial attacks in autonomous vehicle technology. In 2020 IEEE Applied Imagery Pattern Recognition Workshop (pp. 1–7). doi: 10.1109/AIPR50011.2020.9425267.

[43] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1), 18. doi: 10.3390/e23010018.

[44] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft coco: Common objects in context. In European Conference on Computer Vision (pp. 740–755). doi: 10.1007/978-3-319-10602-1_48.

[45] Liu, S., Wang, Q., & Luo, Y. (2019). A review of applications of visual inspection technology based on image processing in the railway industry. *Transportation Safety and Environment*, 1(3), 185–204. doi: 10.1093/tse/tdz007.

[46] Marrone, S., Olivieri, S., Piantadosi, G., & Sansone, C. (2019). Reproducibility of deep CNN for biomedical image processing across frameworks and architectures. In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1–5). doi: 10.23919/EUSIPCO.2019.8902690.

[47] Milburn, D., & Erskine, M. (2019). Digital train control: functional safety for AI based systems. International Railway Safety Council Conference 2019, Perth, Australia.

[48] MOVINGRAIL (2020). Deliverable D3.3: Proposals for Virtual Coupling Communication Structures. https://movingrail.eu/images/Deliverables/D3.3-MOVINGRAIL---Proposals-for-Virtual-Coupling-Communication-Structures.pdf. Accessed 7 October 2022.

[49] Nativi, S., & De Nigris, S. (2021). AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework. EUR 30772 EN. Publications Office of the European Union, Luxembourg. doi: 10.2760/376602.

[50] Network Rail (2022). Public performance measure and delay responsibility. https://www.networkrail.co.uk/who-we-are/how-we-work/performance/railway-performance/public-performance-measure-and-delay-responsibility/. Accessed 9 December 2022.

[51] Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2021). Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(8), 10142–10162. doi: 10.1109/TITS.2021.3122865.

[52] Orive, D., Iriondo, N., Burgos, A., Saráchaga, I., Álvarez, M.L., & Marcos, M. (2019). Fault injection in digital twin as a means to test the response to process faults at virtual commissioning. In 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (pp. 1230–1234). doi: 10.1109/ETFA.2019.8869334.

[53] Pan, S.J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359. doi: 10.1109/TKDE.2009.191.

[54] Pappaterra, M.J. (2022). A literature review for the application of artificial intelligence in the maintenance of railway operations with an emphasis on data. In: Dependable Computing – EDCC 2022 Workshops. EDCC 2022. Communications in Computer and Information Science, 1656. Cham: Springer. doi: 10.1007/978-3-031-16245-9_5.

[55] Pappaterra, M.J., Flammini, F., Vittorini, V., & Bešinović, N. (2021). A systematic review of artificial intelligence public datasets for railway applications. *Infrastructures*, 6(10), 136. doi: 10.3390/infrastructures6100136.

[56] Quinlan, J.R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3), 221 234.

[57] RAILS (2019). Roadmaps for AI integration in the raiL Sector (2019-2023). https://rails-project.eu. Accessed 3 October 2022.

[58] RAILS (2021a). Deliverable 1.2: Summary of existing relevant projects and state-of-the-art of AI applications in railways. RAILS: Roadmaps for AI integration in the raiL Sector. https://rails-project.eu/downloads/deliverables/. Accessed 3 October 2022.

[59] RAILS (2021b). Deliverable D1.3: Application Areas. RAILS: Roadmaps for AI integration in the raiL Sector. https://rails-project.eu/downloads/deliverables/. Accessed 3 October 2022.

[60] RAILS (2022a). Deliverable D2.1: WP2 Report on case studies and analysis of transferability from other sectors. RAILS: Roadmaps for AI integration in the raiL Sector. https://rails-project.eu/downloads/deliverables/. Accessed 3 October 2022.

[61] RAILS (2022b). Deliverable D 3.1: WP3 Report on case studies and analysis of transferability from other sectors. RAILS: Roadmaps for AI integration in the raiL Sector. https://rails-project.eu/downloads/deliverables/. Accessed 3 October 2022.

[62] RAILS (2022c). Deliverable D 4.1: WP4 Report on case studies and analysis of transferability from other sectors. RAILS: Roadmaps for AI integration in the raiL Sector. https://rails-project.eu/downloads/deliverables/. Accessed 3 October 2022.

[63] RAILS (2022d). Deliverable D 3.2: WP3 Report on AI approaches and models. RAILS: Roadmaps for AI integration in the raiL Sector. https://rails-project.eu/downloads/deliverables/. Accessed 3 October 2022.

[64] RAILS (2022e). Deliverable D 2.2: WP2 Report on AI approaches and models. RAILS: Roadmaps for AI integration in the raiL Sector. https://rails-project.eu/downloads/deliverables/. Accessed 3 October 2022.

[65] Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering*, 6(3), 346–360. doi: 10.1016/j.eng.2019.12.012.
[66] Ristić-Durrant, D., Franke, M., Michels, K., Nikolić, V., Banić, M., & Simonović, M. (2021b). Deep learning-based obstacle detection and distance estimation using object bounding box. Facta Universitatis. *Series: Automatic Control and Robotics*, 20(2), 075–085. doi: 10.22190/FUACR210319006R.
[67] Ristić-Durrant, D., Franke, M., & Michles, K. (2021a). A review of vision-based on-board obstacle detection and distance estimation in railways. *Sensors*, 21(10), 3452. doi: 10.3390/s21103452.
[68] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234–241). Cham: Springer.
[69] SAFER-LC (2019). SAFER Level Crossing by integrating and optimizing road-rail infrastructure management and design (2017-2020). https://safer-lc.eu. Accessed 6 October 2022.
[70] Singh, P., Dulebenets, M.A., Pasha, J., Gonzalez, E.D.S., Lau, Y.-Y., & Kampmann, R. (2021). Deployment of autonomous trains in rail transportation: Current trends and existing challenges. IEEE Access, 9, 91427-91461. doi: 10.1109/ACCESS.2021.3091550.
[71] SMART (2016). Smart Automation of Rail Transport (2016-2019). http://smart.masfak.ni.ac.rs. Accessed 6 October 2022.
[72] SMART2 (2019). Advanced integrated obstacle and track intrusion detection system for smart automation of rail transport (2019-2022). https://smart2rail-project.net. Accessed 6 October 2022.
[73] Syeda, K.N., Shirazi, S.N., Naqvi, S.A.A., Parkinson, H.J., & Bamford, G. (2019). Big data and natural language processing for analysing railway safety: Analysis of railway incident reports. In Human Performance Technology: Concepts, Methodologies, Tools, and Applications (pp. 781–809). IGI Global.
[74] Tang, R., De Donato, L., Beš inović, N., Flammini, F., Goverde, R.M.P., Lin, Z., Liu, R., Tang, T., Vittorini, V., Wang, Z. (2022). A literature review of Artificial Intelligence applications in railway systems. *Transportation Research Part C: Emerging Technologies*, 140, 103679. doi: 10.1016/j.trc.2022.103679.
[75] Turing, A.M. (2009). Computing machinery and intelligence. In *Parsing the Turing test* (pp. 23–65). Dordrecht: Springer.
[76] Wang, D., Cui, P., & Zhu, W. (2016). Structural deep network embedding. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1225–1234).
[77] Wang, J., Shi, L., Zhao, Y., Zhang, H., & Szczerbicki, E. (2022). Adversarial attack algorithm for traffic sign recognition. *Multimedia Tools and Applications*, 1–13. doi:10.1007/s11042-022-14067-5.
[78] X2RAIL-1 (2016). Start-up activities for Advanced Signalling and Automation Systems (2016-2021). https://projects.shift2rail.org/s2r_ip2_n.aspx?p=X2RAIL-1. Accessed 7 October 2022.
[79] X2RAIL-3 (2018). Advanced Signalling, Automation and Communication System (IP2 and IP5) – Prototyping the future by means of capacity increase, autonomy and flexible communication (2018-2021). https://projects.shift2rail.org/s2r_ip2_n.aspx?p=X2RAIL-3. Accessed 7 October 2022.
[80] X2RAIL-4 (2019). Advanced signalling and automation system - Completion of activities for enhanced automation systems, train integrity, traffic management evolution and smart object controllers (2019-2023). https://projects.shift2rail.org/s2r_ip2_n.aspx?p=X2RAIL-4. Accessed 7 October 2022.
[81] Xie, P., Li, T., Liu, J., Du, S., Yang, X., & Zhang, J. (2020). Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion*, 59, 1–12. doi: 10.1016/j.inffus.2020.01.002.
[82] Yang, D., Ma, Z., & Buja, A. (2011). A sparse SVD method for high-dimensional data. arXiv preprint arXiv:1112.2433.
[83] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology* (TIST), 10(2), 1–19. doi: 10.1145/3298981.
[84] Yang, X., Liu, W., Zhang, S., Liu, W., & Tao, D. (2020). Targeted attention attack on deep learning models in road sign recognition. *IEEE Internet of Things Journal*, 8(6), 4980–4990. doi: 10.1109/JIOT.2020.3034899.

[85] Yin, M., Li, K., & Cheng, X. (2020). A review on artificial intelligence in high-speed rail. *Transportation Safety and Environment*, 2(4), 247–259. doi: 10.1093/tse/tdaa022
[86] Zendel, O., Murschitz, M., Zeilinger, M., Steininger, D., Abbasi, S., & Beleznai, C. (2019). Railsem19: A dataset for semantic rail scene understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
[87] Zhong, Y., Liu, X., Zhai, D., Jiang, J., & Ji, X. (2022). Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15345–15354).
[88] Zou, X. (2019). A review of object detection techniques. In 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA) (pp. 251–254). doi: 10.1109/ICSGEA.2019.00065.