UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

Learning from accidents

Alawad, Hamad Ali H; Kaewunruen, Sakdirat; An, Min

DOI:
10.1109/ACCESS.2019.2962072

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Alawad, HAH, Kaewunruen, S & An, M 2019, 'Learning from accidents: machine learning for safety at railway stations ', *IEEE Access*, vol. 8, pp. 633-648. https://doi.org/10.1109/ACCESS.2019.2962072

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.
•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
•Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 19. Apr. 2024

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 5, 2019, accepted December 19, 2019, date of publication December 24, 2019, date of current version January 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962072

Learning From Accidents: Machine Learning for Safety at Railway Stations

HAMAD ALAWAD[1], SAKDIRAT KAEWUNRUEN[1], AND MIN AN[2]
[1]Birmingham Centre for Railway Research and Education, University of Birmingham, Birmingham B15 2TT, U.K.
[2]School of Built Environment, University of Salford, Manchester M5 4WT, U.K.

Corresponding author: Sakdirat Kaewunruen (s.kaewunruen@bham.ac.uk)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme through Marie Skłodowska-Curie under Grant 691135, and in part by the APC through the University of Birmingham Library's Open Access Fund. The work of S. Kaewunruen was supported in part by the Australian Academy of Science, in part by the Japan Society for the Promotion of Science for his Invitation Research Fellowship (Long-Term), under Grant JSPS-L15701, in part by the Railway Technical Research Institute, in part by The University of Tokyo, Japan, and in part by the European Commission through the H2020-RISE Project under Grant 691135 Rail Infrastructure Systems Engineering Network (RISEN), which enables a global research network that addresses the grand challenge of railway infrastructure resilience and advanced sensing in extreme environments (www.risen2rail.eu) [70].

ABSTRACT In railway systems, station safety is a critical aspect of the overall structure, and yet, accidents at stations still occur. It is time to learn from these errors and improve conventional methods by utilising the latest technology, such as machine learning (ML), to analyse accidents and enhance safety systems. ML has been employed in many fields, including engineering systems, and it interacts with us throughout our daily lives. Thus, we must consider the available technology in general and ML in particular in the context of safety in the railway industry. This paper explores the employment of the decision tree (DT) method in safety classification and the analysis of accidents at railway stations to predict the traits of passengers affected by accidents. The critical contribution of this study is the presentation of ML and an explanation of how this technique is applied for ensuring safety, utilizing automated processes, and gaining benefits from this powerful technology. To apply and explore this method, a case study has been selected that focuses on the fatalities caused by accidents at railway stations. An analysis of some of these fatal accidents as reported by the Rail Safety and Standards Board (RSSB) is performed and presented in this paper to provide a broader summary of the application of supervised ML for improving safety at railway stations. Finally, this research shows the vast potential of the innovative application of ML in safety analysis for the railway industry.

INDEX TERMS Decision tree, machine learning, railway accidents, railway safety, railway station.

I. INTRODUCTION

The growth in technology has expanded into a vast variety of systems, methodologies, and tools for developing policies in society. There is now a demand to implement artificial intelligence (AI) to interpret the 21st century's ever-growing difficulties in nearly every industry and to focus on promoting intelligent systems interactively. Many of these aspects call for a move towards greater intelligence and a greater sharing of data [1]. Industrial organisations are racing into the AI domain, which is being used to improve safety, analytics and accessibility, and real-time intelligent scheduling, thereby increasing productivity. Applications of AI can reduce safety incidents through reductions in downtime, defects and waste. In self-driving vehicles, for instance, passive safety systems have moved beyond traditional systems towards active ones that are able to monitor their surroundings and can act to prevent collisions and mitigate human failure [2]. The main concern for condition monitoring is the translation of data into information and subsequent employment of that information to improve processes. Machine learning (ML) is a technique for discovering information with self-learning techniques [3], and it has been used in every field due to its ability to obtain useful information from large sets of data [4]. The sector responsible for the railways in the UK, for example, has strategies for digitalising the industry.

There is an opportunity for digital technologies to grant improved levels of safety, in addition to reducing the risk of possible harm to passengers and rail operators. Increasing demand and the capacity of rail networks are important challenges, meaning that potential overcrowding and sometimes delays at peak times are familiar scenes at railway stations.

The associate editor coordinating the review of this manuscript and approving it for publication was Omid Kavehei.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

Incidents are often responsible for delays, and the impact of such events continues to increase. Some older rail stations were designed for closed environments, narrow scopes, and high personnel and facility densities; if an emergency or hazard occurs, there is an expectation of considerable individual harm and loss of assets. Thus, the safety of stations and technology can be used to recognise any deficiency of those stations [5]. New technologies, such as ML, present an opportunity to address these concerns [6]. Moreover, this modernisation may have many direct and indirect impacts, such as national economic growth, and other benefits such as improved safety for passengers and workers, reduced costs, greater sustainability for assets increased service quality and reliability, and improved operation and maintenance [7]. In this study, we apply a decision tree method to examine how accident information or safety records (i.e., age, day, time, gender, and accident category) assist in decisions, enhance the development of loss prevention strategies in the industry and improve safety in railway stations. This paper is divided into seven parts: I. The introduction, II. The contribution, III. The related work on decision trees, IV. Railway station safety and ML, V. The case study, VI. The discussion, and finally, VII. The conclusion.

II. THE CONTRIBUTION

Diekmann [8] indicated that modern methods were emerging and would be able to analyse complex risks. Some of this progress has become evident in AI and the cognitive sciences. Nevertheless, implementation has not yet been fully realised since Diekmann's [8] prophecy. On the other hand, the application of AI has become more attractive due to the progressive refinement of its models, its reduced cost, and improvement of employees' skills and lifestyle (digitalisation) as well as increases in computing power [9], [10]. This paper utilizes an ML method, the decision tree (DT) method, to show how this technique can enhance both safety and the analysis of accidents and address risk methodology gaps in railway stations. Our main contribution is a method for automatic railway safety classification and analysis through safety records. The history of accidents in UK stations has been investigated. For this process, we designed a different DT using ML classification software. Two labelled datasets with varying types of accidents were constructed from the calibration run accident reports. Furthermore, we propose a framework for railway station safety benefits based on both internal and external safety data and real-time data to enable the construction of smart stations in the future. The principal objective of this study on safety predictions lies in how to apply ML to establish a prediction model and analyse accidents given a more comprehensive understanding of the risks with an acceptable level of accuracy.

III. RELATED WORK

A. RAILWAY APPLICATIONS AND MACHINE LEARNING

This paper reviews an extensive collection of literature examining the use of ML in the railway industry. The findings from the relevant research are provided in the next section. It has been found that railway maintenance is essential and decisive for ensuring safety and quality; however, it is costly from an economic perspective. Thus, the maintenance operations in the railway industry and monitoring have drawn attention by many scholars [11]. We present in this section previous studies that have employed ML in the field of infrastructure, operations and trains or the components of systems, including maintenance and monitoring. The Swedish Transport Administration (Trafikverket) first suggested applying ML analysis in big data technology to maintenance activities, therein aiming for safe and robust railway assets [12]. To predict the conditions that might lead to failures of railway tracks/trains and to improve rail network speeds and railway predictive maintenance, an ML approach has been proposed [13]. Moreover, comparisons of a specialized support vector machine (SVM) with the DT technique have shown a significantly better performance under the customised SVM [13]. Additionally, the classification of image data by a multi-layer perceptron and SVM has been performed to automate the process of visual condition monitoring of wooden railway sleepers, therein achieving high classification accuracy [14]. For railway track beds, an ML classifier method has been proposed for recognising woody plants [15]. For detecting obstacles on the track, utilising ML technology in comparing input and reference data to train frontal view camera pictures was proposed, therein yielding accurate and successful results in experiments [16]. Moreover, to improve the detection of defects in railway fasteners for improving accuracy and overall safety, ML has been applied to image recognition on railway tracks [3], [17]. Furthermore, to classify wheel failures, a logistic regression model has been developed to predict the possibility of events of high wheel effect train stops, where the results also showed high accuracy [18]. During normal operating speeds and for defect detection in railway train wheels, a sensor system on a railway network has been developed for vertical force wheel measurements. Two ML methods have performed classification with SVM and artificial neural networks for image classification. The modes analyse multiple time series of the vertical force of a wheel to determine whether a wheel has a defect [19]. For high-speed train tracks, the data from maintenance records have been utilised to predict faults, where the results reveal that the support vector regression outperforms other employed techniques [11]. Track geometry conditions have been selected for maintenance; thus, supervised and unsupervised ML methods are applied to big data to predict the effects of geocell installation on the track geometry quality. For Dutch railway tracks, operators have been using big data methods to facilitate maintenance decision making, which has shown great potential for railway track condition monitoring [20].

Additionally, to assess the risk of a rail failure on the tracks of the Dutch railway network, a big data analysis approach has been used, with a large number of records from video cameras as input [21]. Big data technology has been presented for improving decision making for marketing

decision makers of railway freight [22]. A survey covering operations, maintenance and safety was conducted to provide a comprehensive review of the applications of big data for the railway [23]. Supervised ML techniques achieve the lowest prediction error and can learn and classify defective tracks from non-defective sections [24]. For railway passenger volume forecasting, SVM optimised by a genetic algorithm (GA-SVM) has been applied to prediction approaches for passenger volumes for railways in China. This method has achieved greater forecasting accuracy compared with artificial neural networks [25]. For timetable improvement and real-time delay monitoring in a range of real train networks of the Deutsche Bahn, a delay prediction system has been developed utilising a neural network [26]. For studying and analysing large volumes of data, ML methods are growing increasingly powerful for track condition prediction, therein achieving improvements in future railway safety and service quality [27]–[29]. A vision-based object detection algorithm for passenger safety on a railway platform that detects risks in stations in real time has been proposed [30], [31]. Some additional related work is presented in Table 1. In conclusion, the related work discussed above presents a range of approaches taken for researching ML in the railway industry and how such advanced technology is being utilised to advance the big data revolution in the context of the railway industry.

TABLE 1. Examples of studies utilising advanced methods in railway applications.

Year/ References	The technique	The data source	More details
2017, [24]	Machine learning	Using a track geometry car	Track geometry
2017, [17]	Deep learning, CNN	Collected by track inspection vehicle	Automated track inspection
2017, [32]	Machine learning	From library and a database of historical incidents	Estimates the probability of the failure of network operation
2018, [33]	Deep learning, Fuzzy inference model	From video cameras	A decision support approach for rail maintenance
2018, [34]	Deep learning, R-CNN	Images by inspection vehicle	Automatic visual inspection (object detection)
2018, [35]	Machine learning	From train-borne system	optical fibre sensing for track defects
2018, [36]	Deep learning, text mining	Historical accidents report	classify accident causes
2018, [19]	Machine learning' SVM and CNN	installed sensor system on the railway network	dynamics Wheel defects on railway
2018, [37]	Machine learning, ANN	simulation-based dataset of signal response by the alternating current field measurement (ACFM) sensor	fatigue crack detection and sizing
2019, [38]	Deep learning, Kernel principal component analysis SVM	Laser ultrasonic	classification for rail defects
2019, [39]	Bayesian network	historical inspection data	monitoring of railway catenaries for railway infrastructure maintenance

B. ML AND DTS BACKGROUND

This section introduces ML and supervised learning, which are related to our paper. ML is particularly important in DT, and a brief description is given below. ML models propose to ''learn'' the association between a set of input and output data. Scholars engaged in AI desire to explore whether machines can learn from historical data to produce reliable decisions and conclusions, and the field of ML has obtained substantial momentum. Improvements in computing and communications technologies have led to a strengthening of the argument for applying complicated numerical predictions to big data, as it would become increasingly fast over time. Some examples of sectors applying ML are the following:

- Financial (assessing risk and fraud detection)
- Healthcare (diagnostic care and health monitoring)
- Retail (Online recommendations and marketing)

There are three main types of ML. One type is supervised learning, which requires labelled data to train models and make predictions. The second type is unsupervised learning, which determines patterns from unlabelled data. The third type is reinforcement learning, which enhances learning from feedback obtained from interactions with external environments. Numerous classifier models have been used in several fields, and each model has benefits and limitations in performing experiments based on research needs. Linear discriminant analysis (LDA) and naive Bayes provide probabilities, and samples belonging to classes of SVM and neural networks perform better on multidimensional and continuous feature datasets. The k-nearest neighbour method is sensitive to irrelevant data and intolerant of noise. The naive Bayes classifier is fast because it requires minimal storage. The DT model has an important interpretability for promoting further analysis of the dataset [4]. We assume supervised learning in this work and that the classification process implements DT based on ML software [40]. Additionally, a review of classification techniques with supervised learning algorithms is given in the literature [41]. DTs are trees that group instances by classifying them based on background values. The objective is to build a model capable of predicting the value of a target variable by learning simple decision rules understood from the data features. Each node in a DT draws an element in a case to be organised, and each branch

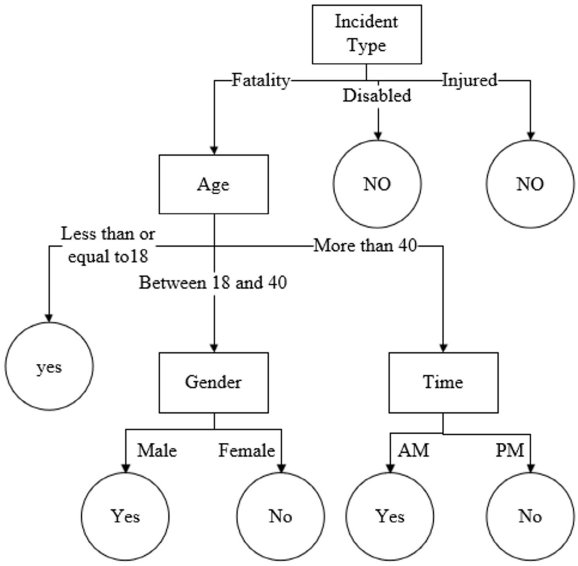

FIGURE 1. Examples of DT for the incident type training set.

TABLE 2. Training set of incident type.

Incident Type	Age	Gender	Time	Class
Fatality	Less than or equal to 18	Male	AM	Yes
Fatality	Less than or equal to 18	Male	PM	Yes
Fatality	Between 18 and 40	Male	AM	Yes
Fatality	Between 18 and 40	Female	PM	No
Fatality	more than or equal to 40	Male	AM	Yes
Fatality	More than or equal to 40	Male	PM	No
Disabled	Between 18 and 40	Female	PM	No
Injured	Between 18 and 40	Female	PM	No

describes a value that the node can find. Fig. 1 is an instance of a DT for the training set of Table 2. There are several variants of DTs such as classification and regression trees (CART), chi-squared automatic interaction detection (CHAID), and iterative dichotomiser ID3, C4.5, and C5.0 [42]–[44].

Using the DT described in Fig. 1 as an example, the instance (incident type, age, gender, and time) will be used to classify the nodes as incident type, age, and gender, which would categorise the instance as being positive (classified as Yes).

DT algorithms use a set of supervised learned decision rules for predictions based on inputs of selected predictor factors and learning from overlapping attributes; moreover, it has been shown that the DTs have satisfactory computational performance and easier logical explanations. The model is based on the DT model based on CART. The algorithm in the software that was utilised in the model was inspired by Breiman's [45] CART DT models 1984. CART is a DT algorithm that produces binary classification or regression trees, depending on whether the target variable is categorical or numeric, and extracts the existing patterns or rules found in the dataset. The model with CART is substantially more scalable and able to address multiple data types simultaneously. The model stops growing when they have exhausted their ability to better fit the training data. Each tree node attempts to split the data in the most optimal manner so that the classification splits maximize the information gain.

IV. RAILWAY STATION SAFETY AND ML

Stations, as a dynamic environment, require a dynamic operation and safety process that reflects the nature of risks. Thus, a novel dynamic method must increase the safety and support decision makers in a timely manner [46], [47]. Moreover, there are several drawbacks of conventional methods that need to be mitigated, e.g., uncertainty [48] and safety information, and the risk plan outcomes are sometimes based on values from several decades ago [49]. Another drawback is that traditional static analysis is too static and not regularly updated, thereby being unable to capture the changes in the process and plan [50], [51]. The drawbacks of the traditional methods of risk assessment need improvement under dynamic risk [52]. Passenger safety, security and risk management are the primary goals of railway systems, and managing and enhancing safety and ensuring reliable environments within a railway station are one of the most significant challenges. The stations contain physical objects, people, and multiple systems (e.g., closed-circuit television (CCTV), heating, ventilation, and air conditioning (HVAC), fire systems, and screening systems). Various accidents, such as passengers falling from the platform or being caught between train doors, electrical shocks, slipping/tripping incidents, vandalism and fire, have occurred at stations. The complexity of the stations, their dynamic nature and safety challenges have demonstrated the need for intelligent dynamic automatic technology, such as ML, to mitigate safety challenges and meet future requirements. ML has contributed to the prediction of safety in construction and other construction aspects such as cost, time and quality as well as accident occurrence and severity [53]. The big data revolution is now universally known in the railway industry, and there is a need for the capability to process a growing amount of data; the concept of smart railway stations offers a thriving environment for big data strategies, and smart safety is expected to play an essential role in managing risk and safety at stations. Safety managers of stations use numerous forms, software and data collection to ensure that the station is safe and that every task is compliant with safety and security plans. A smarter safety expression utilising ML and converting data into knowledge has been proposed to further deliver safer stations. Open-source data, sensor technology, and predictive analytics can be used to improve compliance with regulations designed to keep the stations safer. Innovative technologies aid the industry and enhance security and safety at stations. This has increased timetabling, predicted demand and improved decision making through data processing [31], [54]. Thus, the power of computers and the capabilities of ML for training

can be used for analysing accidents and assessing the risks facing safety-critical infrastructure such as railways. This would allow for the processing of big data in the form of indicators from daily operations and from historical data accidents, which would be used for training and testing the model and then implementing a reliable, robust model for facilitating real-time safety monitoring in railway stations.

A. SAFETY AND ML (APPROXIMATION MODEL)

The objective is to minimise the risk, which is an important aspect of ML. In this section, we present the functional estimation of the model, which makes it implicit that risk is a functional R(m). It is suggested that the learning steps can be divided into three stages:

1- A random vector x that is captured independently from a fixed but unknown distribution $P(x)$ must be generated.
2- The output vector is assumed as y, which is returned by the supervisor for every input vector according to a conditional distribution function $P(y \mid x)$, which is also fixed but unknown.
3- The learning machine is able to execute a set of functions $f(x, w), w \in W$. The best scenario of the response or the supervisor's response is selected as a step in the ML process from the given set of functions based on a training set of t independent observations:

$$(x1, y1), (x2, y2)(xt,yt) \tag{1}$$

This shows that learning corresponds to the problem of function estimation. To find the risk functional, R(m), we need to consider the loss or discrepancy L(y, z), where y is the response of the supervisor to a given input x and z is the response functional provided by the learning machine, where z = f (x, w) (see part three of the learning steps) and the loss will be L(y, f (x, w)). Thus, the expected value of the discrepancy, given by the risk functional, is

$$R(m) = \int L(y, f(x, w))dP(x \mid y) \tag{2}$$

Over the set of functions $f(x, w), w \in W$, the target is to minimise the risk functional R (m). However, the joint probability distribution $P(x \mid y) = P(y \mid x)\, P(x)$ is unknown, and the only available information is contained in the training set (1) [55]. The risk minimisation approach to ML has shown strengths in practical applications and has the ability to capture the safety risk component.

However, it does not capture issues related to uncertainty and loss functions that are relevant for safety. To enhance safety with ML, four groups of principles have been classified:

- Safety reserves (safety factors and margins)
- Inherently safe design (replace dangerous material by less dangerous materials)
- Safe failure (system remains safe when it fails)
- Procedure safeguards (training, quality, standards, etc.)

To extend the ML model beyond risk reduction for improving safety, it has been suggested that each of these principles should be sought [56].

B. FLOWCHART OF ML IN THE SAFETY PROCESS

Given that railway stations are crowded areas and pose a challenge to safety and security, efforts do not fall exclusively to the state or the stockholders but rather relate to society as a whole. The stations have certain characteristics, such as being crowded and complicated, and may have weak management systems. Many systems located in the stations, with their open structure, characterize the complexity of the railway stations. The control and prevention of unexpected events in the stations are critical, and thus, new technology needs to be used more frequently to make them secure and safe. Therefore, railway station system features will be analysed, with the aim of providing suggestions for the improvement and employment of technology and for designing a safety and security framework. There are variants of applications utilising AI technologies such as ML and big-data in many industries, such as medical, banking, and marketing; however, few technologies are being used in railways and transportation. Information on safety, security monitoring and emergency rescue by supervision in the rail system has not yet been entirely generated due to a lack of integrated systems for rail transportation safety and security, as well as delayed implementations of technology. Therefore, there is a lack of ability to utilize significant amounts of data in the railway industry to explain the relationships between operational factors, safety and security, especially in railway stations. Thus, more research is required to validate this relationship, which is the goal of this research, and to design benchmarks for the expected level of safety and security performance in railway stations. In addition, in future work, the obtained data will allow for its validity to be evaluated in a case study of the proposed framework. Massive dataset resources are captured for analysis, including the history of accidents locally and internationally. The concept of a smart city and smart stations represents an advanced level of this technology. Intelligently gathered information, weather conditions and crime can be associated in real time with the railway data centre and used to predict scenarios and consequences. This knowledge discovery from the predictive model will actively aid decision makers, save time and enhance safety and security at stations, therein expecting to obtain high-performance predictions (see Fig. 2).

In the station, there is a range of sources that can be used to find the factors that may form an anchor. First, historical incidents, such as fatal accidents that have been analysed in this study, were chosen. The railway station analysis was selected, which covers many aspects of the railway industry and presents a considerable amount of data. This analysis has shown that extensive data from the railway industry and stations, in particular, can be utilised to implement new technology such as ML. Then, from all the overlapping systems in the stations and the history of incidents, the factors that may

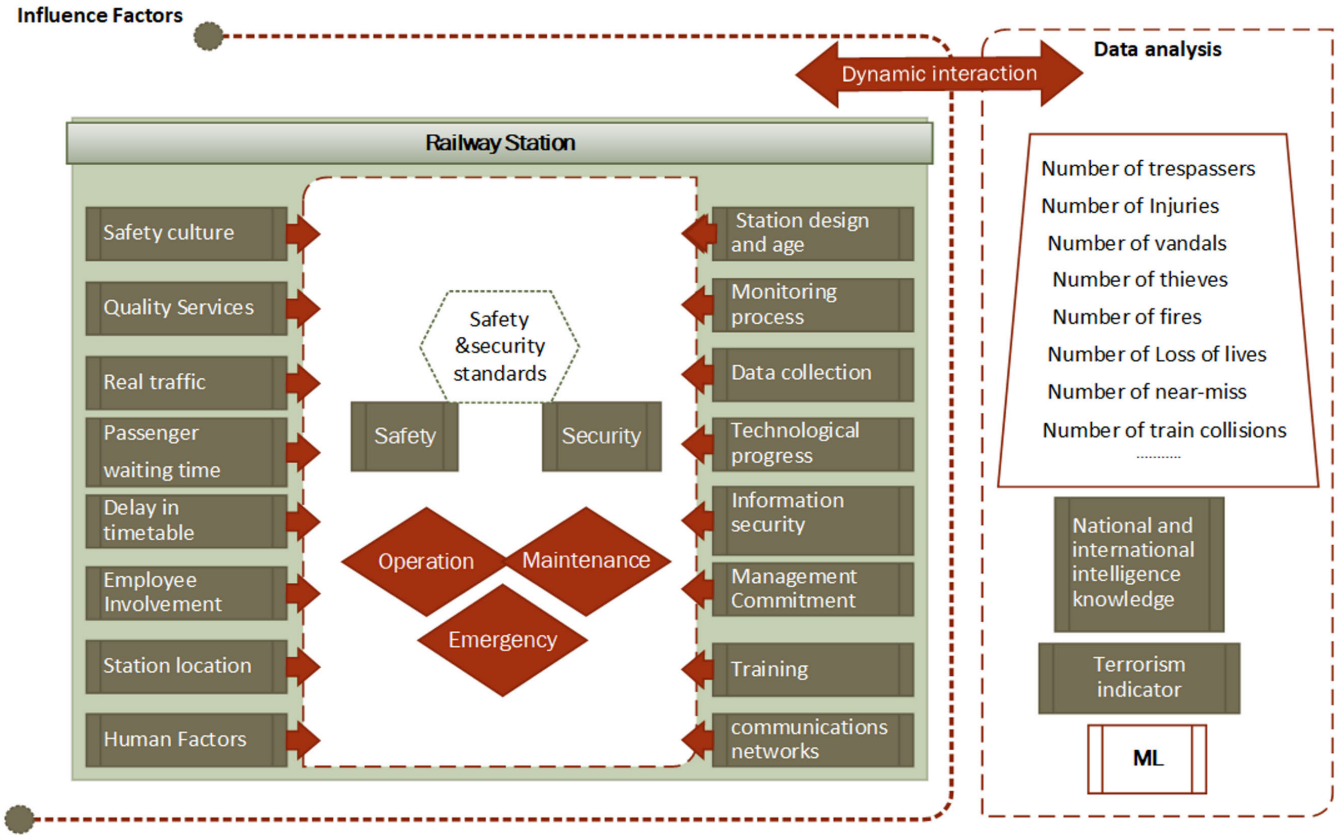

FIGURE 2. Flowchart of proposed ML used in the safety process.

directly or indirectly affect the station's safety and security can be discovered. These factors work as indicators to ensure more effective safety systems in the stations. Moreover, the model aims at advancing measures and supplying an essential basis of absolute safety and security systems, as well as developing safety management and a foundation for a comprehensive design framework including new technologies [57], [58].

V. CASE STUDY

This paper selects a representative sample of accidents that occurred in the stations and lead to fatalities. The aim of this case study is to expound upon the potential for applying supervised ML to the railway industry. The importance of this study is in its explanation of the potential of ML to be used in improving services, management and, in particular, safety in station environments. This designed model for predicting safety and supporting decision makers is based on data collected from rail reports (RSSB) since 2002. The collected data on accidents that have been reported and published represent a selection of 80 incidents at stations in the UK that have been or are subject to an investigation by the UK's national investigation body: the Rail Accident Investigation Branch (RAIB) [59], [60]. The process of applying supervised ML is a process of learning a set of rules from instances (examples in a training set). Generally, the first step in the supervised learning method is collecting the dataset and finding the attributes that are the most informative. The next step is preparing the data; in most cases, the data contain noise and missing feature values and consequently require meaningful pre-processing [17], [61]. Next, the classifier model is selected, and to calculate a classifier's accuracy, we split the dataset for training the model and evaluations.

A. DATA PREPARATION

Data preparation is a fundamental stage of data analysis. Data pre-processing consumes more than 60% of the total effort in the modelling process on average; this is important because of the impact on the results. The limited availability of data is challenging for many researchers, in particular, who utilise AI methods that need massive amounts of data to gain the benefits of such technology. In this work, data that only satisfy the conditions have been collected. Accidents also lead to deaths within the station's boundaries; this gives the research greater precision and indicates importance for the worst-case scenarios. This work has relied on trusted sources, such as investigation reports, and it has excluded other sources that may not provide all the specifics of the accidents. The data that did provide information on the passengers or the details of the accident were omitted to ensure that there were no missing attributes. The number of accidents was 80 (see Fig. 3).

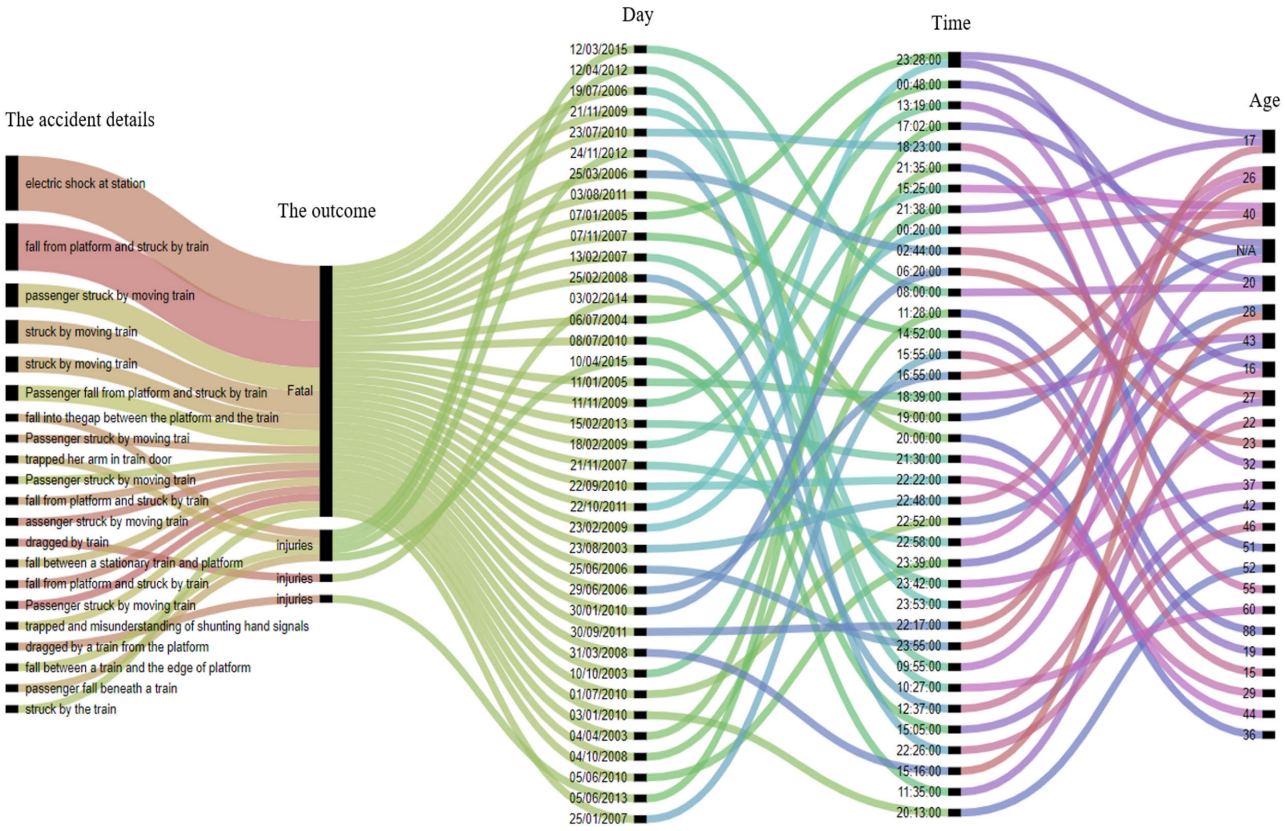

FIGURE 3. Example of an alluvial diagram of raw accident data: accident details, outcome, day, time, and passenger age.

Some operations have been conducted to modify the data structure to fit the modelling process, including the following:

- Generalization: For example, the date of the accident field in the accident documents, which consists of the year, month, and the day, is amended to contain the specific day of the week (Saturday (D-1) etc.) and particular time such as AM or PM.
- Designing highlights: From the cause of each accident, for example, falling from the platform and being struck by the train (T1-F), electrical shock (T2-E), or being struck by a moving train (T3-S), a distinct feature is created.
- Transforming data: The set of values is consistent with a new set of feature values. For example, the day of the accident, age and gender (Female (G-F) and Male (G-M)) of the person are converted into discrete values.
- Reducing or removing redundant features: Features that are inappropriate for this study, such as accident occurring out of the stations, the accidents not leading to death or the accidents not having details of the person who was involved, are removed or reduced.

The data selection from the published reports provides factors that might characterise the scenarios of events, such as passenger age and gender, as well as the day of the event and the exact time. Details of the accidents have also been considered. Moreover, depending on the RSSB reports, the data that have been used in this report are cut-off from the industry's safety management information system (SMIS). By preparing and cleaning the data during data exploration, the number of accidents is reduced to 71 accidents (instances), with five variables, resulting in fatalities at the train station boundaries (see Fig. 4). Each accident related to stations and the information from railway industries generally do not have many details except for certain reports on the web. Considering existing data and their value, we work with a small database, and we attempt to make these data more useful; thus, non-relevant information is removed [62], [63].

B. THE ANALYSIS AND CLUSTERING

The dataset of 71 accidents is used in this analysis. This dataset contains the attributes of age, sex of the passengers, the day of the week and the time of the event as well as the cause of the deaths. The attribute matrix was applied as the input of the DT model, and the time was targeted as a predictor. The process of analysing and utilising ML as the method proposed in this paper is used to learn from the accidents and thereby benefit from this technology in the field. There are more selections to model and predict utilising other predictors and many options for inputs. Following the data cleaning step, we analyse the data by applying ML analytics software [64]. Some DTs are used in this work for various predictors in the

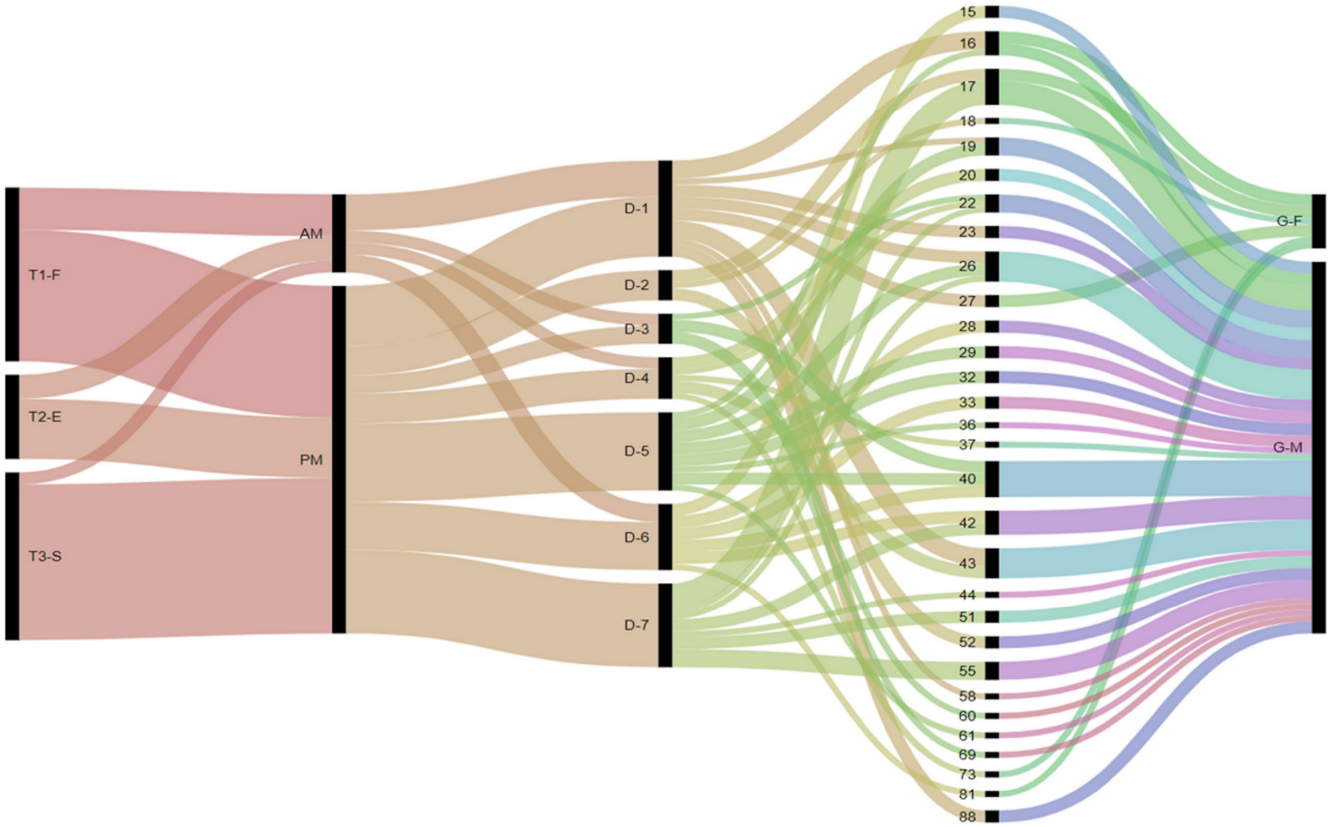

FIGURE 4. Example of an alluvial diagram of accident data after processing, showing the accident type, time, day, passenger age and passenger gender.

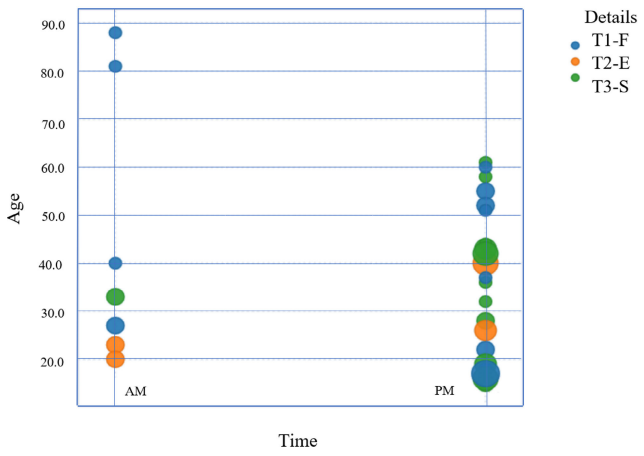

FIGURE 5. Scatterplot showing a correlation between the details of the accidents and time with the passenger age.

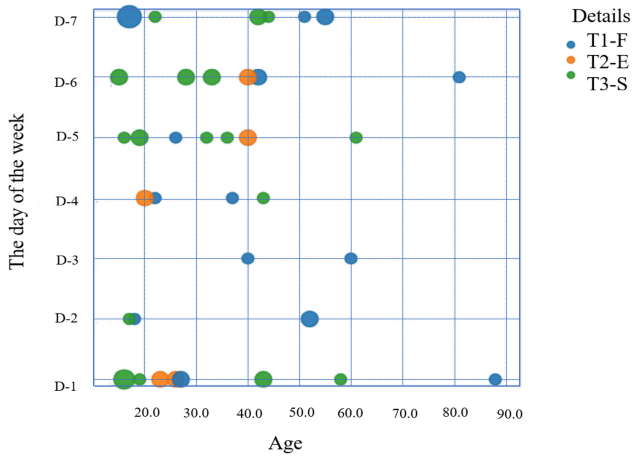

FIGURE 6. Scatterplot showing the correlation between the details of the accidents and day of the week with the passenger age.

method and for proving the power of ML in analysing safety data. The ML tool (MLT) enables us to review and visualise descriptive statistics of the dataset (see Figs. 5 and 6 below).

This also shows the distribution of the passengers' age, time and details for each accident [40]. The DTs in the selected MLT are dynamic methods used to analyse the datasets. Thus, we set the attributes of the accidents as our target; thus, any predictor from the dataset can be used. The accuracy and prediction path will vary from one attribute to another. However, this paper attempts to outline the data to explore the interests for safety data analysis and to demonstrate the suggested method. An example of the results of the DT is shown in Fig. 7. The example indicates how important each factor in the prediction of accidents is, where the day of the week is the most important factor, followed by age (see Table 3).

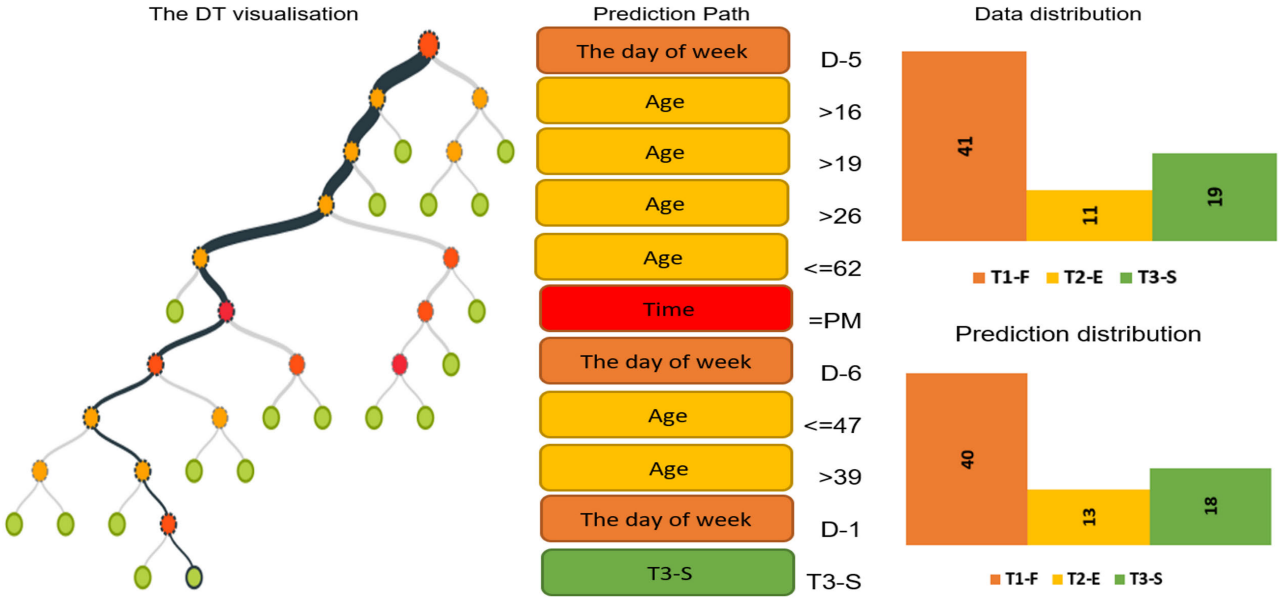

FIGURE 7. DT diagram and graphs showing the prediction and data distribution from the training dataset.

TABLE 3. The analysis of the dataset.

The input factors (attributes)	The analysis	The importance of the field %
Day of the week	Saturday=D1, Sunday=D2 etc.	52.8
Passenger details	Mean age =36.14 (child, elderly, disabled, teenager male or female (G-M, F)	41.41
The details of the accidents	T1-F: Fall from platform and struck T2-S: Struck by moving train T3-E: Electrical shock in the station	3.35
Time	AM, PM	2.48

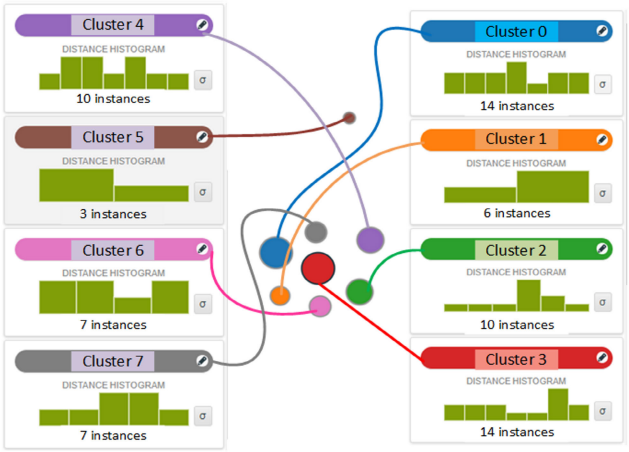

FIGURE 8. The 8-cluster diagram nodes and histograms.

For further exposition of ML techniques in such cases, a clustering method is applied because intelligent methods used to present and extract data patterns of interest are searched, and it is shown that ML is a powerful analysis method for safety and risk management in railway stations. To analyse the unsupervised dataset, ML is chosen with the K-means algorithm (canonical clustering), where the number of clusters is eight. However, the remainder of the work is supervised ML. Utilising cluster analysis involves separating datasets into subsets of instances (clusters) and finding similarities (see Fig. 8 and Table 4). The clusters are placed closer to one another if they are more similar and farther away if they are very dissimilar, where, for example, cluster number 5 is a long distance from the other clusters. The size of a cluster, presented as a circle, is proportional to the number of instances in that group; thus, the largest cluster has 14 cases, and the smallest cluster has only 3 cases. Cluster analysis is often an iterative process that requires some trial and error until the most useful grouping of data instances is achieved. However, we utilise the 8 clusters as the default. To process the learning data, the K-means algorithm from data mining starts with the first group of randomly selected centroids, which are used as the initial points for every cluster, and then performs iterative (repetitive) calculations to optimise the positions of the centroids. The MLT utilises optimised versions of the K-means algorithm; the user needs to specify the number of clusters in advance, here specified as 8.

C. DECISION TREE CLASSIFICATION

A DT is a determination support tool that applies a tree-like pattern of decisions and their likely outcomes [65]. There are

TABLE 4. The details of the clusters.

Centroid name	Instances	Minimum intercentroid distance	Mean intercentroid distance	Maximum Intercentroid distance	Distance sum squares	Distance standard deviation	Distance sum	Distance median	Distance maximum	Distance minimum	Distance variance	Distance mean
Cluster 0	14.00	0.34	0.51	0.98	0.71	0.08	2.96	0.25	0.29	0.08	0.01	0.21
Cluster 1	6.00	0.43	0.54	0.90	0.25	0.13	1.00	0.25	0.25	0.00	0.02	0.17
Cluster 2	10.00	0.36	0.46	0.65	0.57	0.08	2.28	0.25	0.32	0.04	0.01	0.23
Cluster 3	14.00	0.30	0.47	0.89	0.78	0.12	2.87	0.27	0.31	0.01	0.01	0.21
Cluster 4	10.00	0.35	0.52	0.64	0.59	0.11	2.17	0.25	0.41	0.03	0.01	0.22
Cluster 5	3.00	0.53	0.81	0.98	0.13	0.19	0.42	0.03	0.35	0.03	0.03	0.14
Cluster 6	7.00	0.30	0.43	0.73	0.43	0.13	1.51	0.25	0.35	0.04	0.02	0.22
Cluster 7	7.00	0.36	0.54	0.98	0.52	0.04	1.88	0.25	0.35	0.25	0.00	0.27

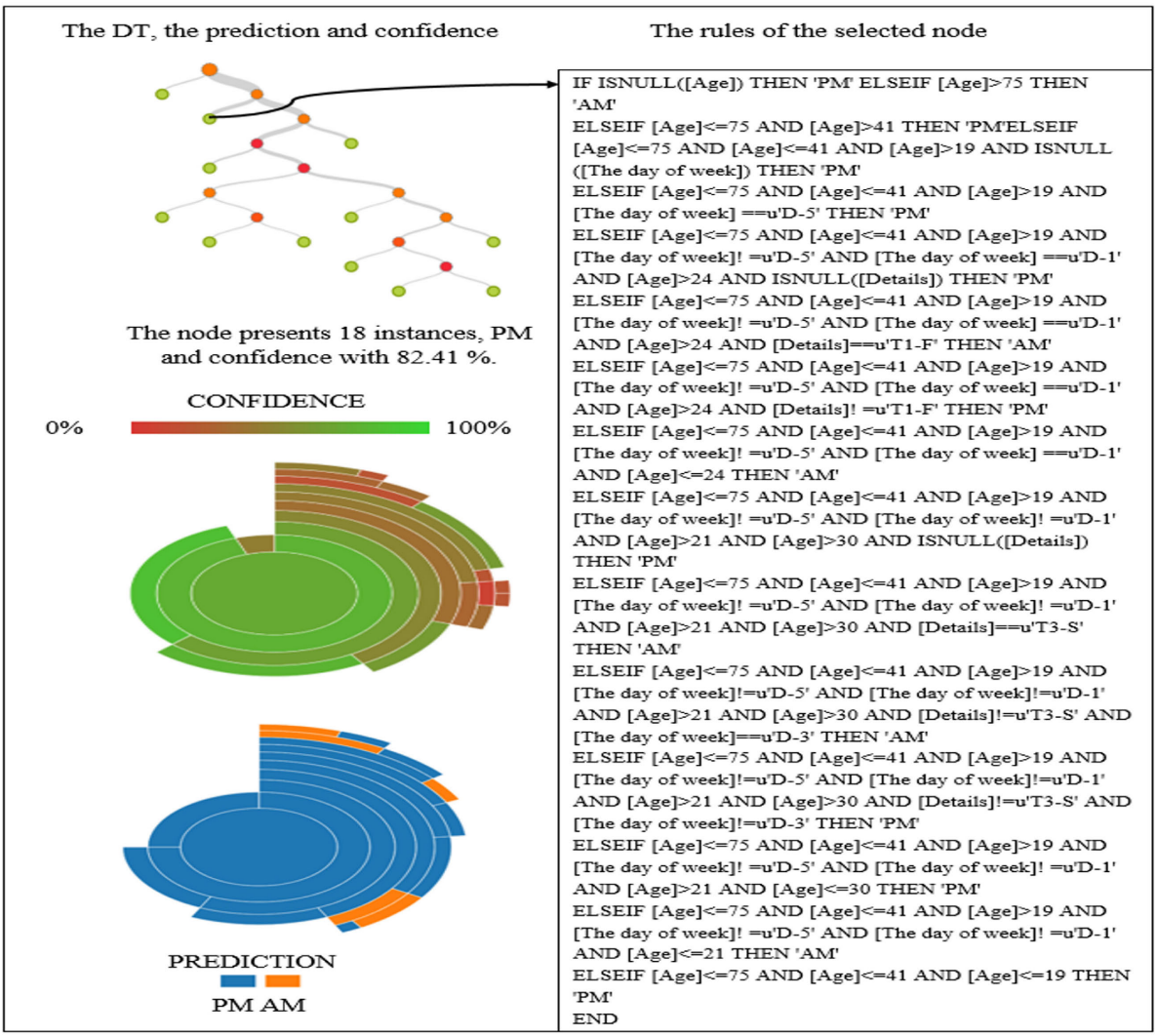

FIGURE 9. The DT with an example of a selected node showing the rules.

many possible ML approaches to safety analysis. In this case, we train a DT to classify the accidents and the patterns that occurred in these accidents in the stations [66]. This model is applied to a wide variety of data, and it is preferable because its structured rules are simple to follow and understand. This technique is used to classify instances by classifying them based on feature values [67]. The two general types of DTs are classification (where the class variable is discrete) and regression (where the class variable is continuous) [67], [68]. After the datasets are uploaded, a DT model is designed and visualised. The DT for the predictive model provides a visualisation of the prediction case. The DTs have useful information, and branches are used to make a branching decision. This shows the decisions that led to a given prediction. The tool presents the model prediction path on the side of the tree, which gives this tool an advantage. The tree has colours that denote the different lists that the branch possesses, which are presented with strengths to classify the predictive path.

D. PREDICTION AND ANALYSIS

The DT model has been applied for predicting the future values of passenger attributes based on previously observed values. In this case study, the target passenger characteristics can be changed from one characteristic to another. This results in a unified framework that can perform analyses of variable data using the ML algorithms. This DT shows the prediction path, where the strength of the path in the tree is indicated by bolded branch paths. The time attribute has been selected as a predictor (see Fig. 9).

Obtaining more details of the prediction path and input data changes of the input fields has been an interesting process. After midday, the prediction shows more accidents for an older passenger at the end of the week. The time represents a critical point as an input field affecting the prediction. There is a slight influence of the day in the forecast, which may refer to other factors not involved in this case study (see Figs. 10 and 11).

Some factors are clear, such as the time, where PM experienced more accidents than AM. The large ratio of accidents occurring for males is seen in two important age groups: the young and the old. Many factors, such as intoxication, must be considered; however, in this research, we attempted to apply ML rather than a deep analysis of the causes of the accident. The DT has been applied, and it shows how the prediction path predicts the target instances. The results present acceptable values that will be further justified with more available data in future research. Depending on the selected data type and the prediction targets, certain algorithms will be chosen. For instance, the accident type T1-F has been targeted to present the numerical data of the prediction path from the DT (see Table 5).

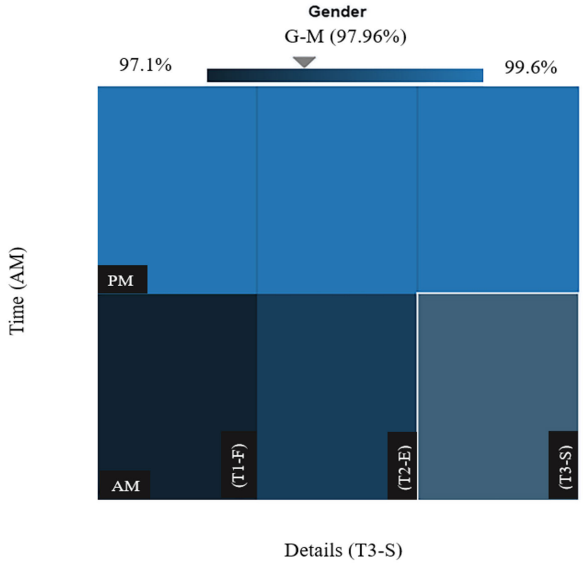

FIGURE 10. Comparisons of the input field on the prediction path with different factors (time and the details of the accidents). The result is shown for the example accident type T3-S during PM.

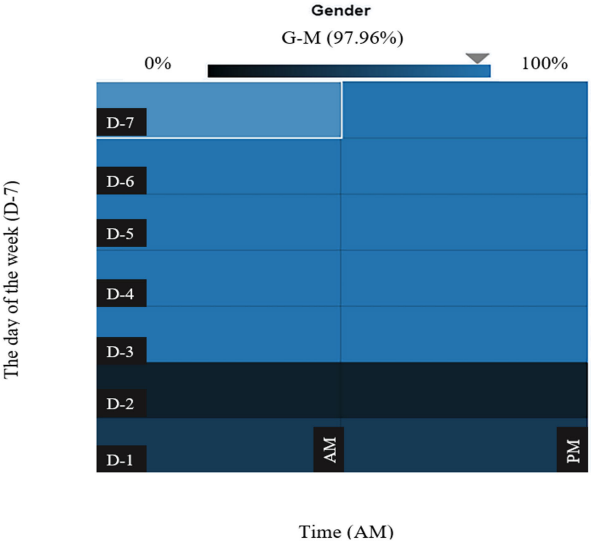

FIGURE 11. Comparisons of input fields on the prediction path with different factors (time and the day of week) for example day 7.

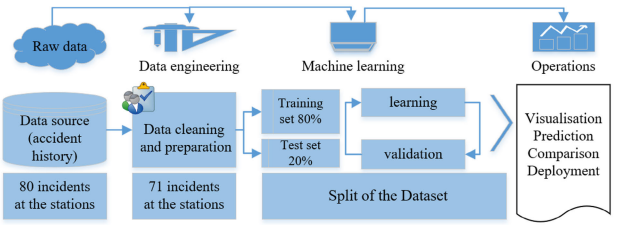

FIGURE 12. The process of splitting the dataset.

E. LEARNING AND VALIDATION

Evaluating the model to ensure that it produces reliable predictions is significant. In this section, we aim to obtain an overview of the model's predictive performance and create a framework for comparing models with different configurations or different algorithms to classify the models with the best predictive performance. The model is built on a subset of the data, termed the training data, and they are applied to

TABLE 5. The DT prediction path example, where the target in this example is the accident type T1-F.

Node Output	Node Instances	Node Support	Node Confidence	Prediction	Prediction Confidence
T1-F	6	0.107	0.609	T1-F	0.609
T1-F	3	0.053	0.207	T3-S	0.609
T3-S	6	0.107	0.609	T3-S	0.609
T1-F	10	0.178	0.595	T1-F	0.595
T2-E	3	0.053	0.207	T3-S	0.510
T3-S	4	0.071	0.510	T3-S	0.510
T1-F	1	0.017	0.206	T2-E	0.510
T2-E	4	0.071	0.510	T2-E	0.510
T1-F	2	0.035	0.342	T3-S	0.452
T3-S	9	0.160	0.452	T3-S	0.452
T3-S	8	0.142	0.409	T3-S	0.409
T1-F	2	0.035	0.342	T2-E	0.375
T2-E	5	0.089	0.375	T2-E	0.375
T3-S	7	0.125	0.358	T3-S	0.358
T2-E	4	0.071	0.510	T3-S	0.353
T3-S	11	0.196	0.353	T3-S	0.353
T1-F	2	0.035	0.094	T1-F	0.342
T1-F	2	0.035	0.342	T1-F	0.342
T1-F	2	0.035	0.342	T3-S	0.342
T3-S	2	0.035	0.342	T3-S	0.342
T2-E	2	0.035	0.342	T1-F	0.342
T1-F	2	0.035	0.342	T1-F	0.342
T1-F	2	0.035	0.342	T1-F	0.342
T1-F	1	0.017	0.206	T3-S	0.342
T3-S	2	0.035	0.342	T3-S	0.342
T3-S	46	0.821	0.302	T3-S	0.302
T1-F	4	0.071	0.300	T1-F	0.300
T1-F	6	0.107	0.299	T1-F	0.299
T1-F	56	1	0.291	T1-F	0.291
T2-E	7	0.125	0.250	T2-E	0.250
T2-E	7	0.125	0.250	T2-E	0.250
T2-E	13	0.232	0.232	T2-E	0.232
T3-S	35	0.625	0.231	T3-S	0.231
T3-S	3	0.053	0.207	T3-S	0.207
T3-S	21	0.375	0.207	T3-S	0.207
T1-F	28	0.5	0.207	T1-F	0.207
T3-S	4	0.071	0.510	T1-F	0.150
T1-F	4	0.071	0.150	T1-F	0.150
T1-F	4	0.071	0.150	T1-F	0.150

predict new data that are not part of this training subset. This useful model has been shown to be well balanced in terms of avoiding both overfitting and underfitting. The MLT extends a training/testing data split by choosing subsets for generating

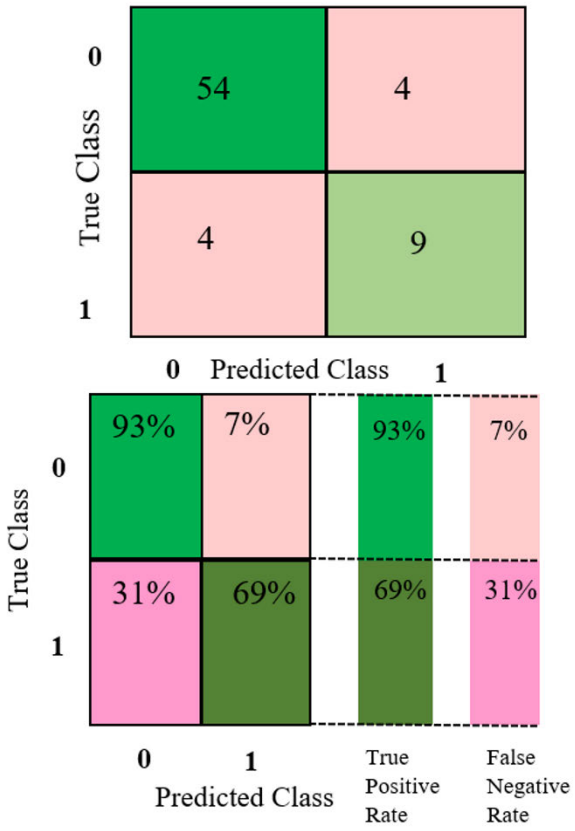

FIGURE 13. The evaluation results of the performance per class in the confusion matrix.

the 80%/20% split of the dataset. The former can be applied to train the model and the latter to test it; thus, 15 accidents are randomly selected for testing, and the remaining 56 are used for training the model (Fig. 12). The accident scenarios in the dataset's matrix include the attributes of the age and sex of the passenger, the day and time, and the cause that led to death.

The MLT provides a way to measure and compare the performance of the models. Moreover, the tool allows for the creation of a new DT model with a modified confidence value. We can now compare and evaluate the DT models. The prediction model is a classifier of the instances between the passenger traits in our prediction model, which depend on the accident parameter history. A two-class prediction was selected (binary classification) in the case of fatal accidents in the stations (to determine whether the accident occurred during PM (0) or AM (1)). The outcomes of the prediction are labelled as either positive or negative. If the prediction is positive and the actual value is also positive, then it is called a true positive (TP); with the same concepts, false positives (FP), true negatives (TN), and false negatives (FN) are realised. The four outcomes can be formulated as a 2 x 2 contingency table or confusion matrix, as shown below (see Fig. 13).

The positive class was chosen as PM in applying this evaluation. Then, some statistical measures, such as accuracy (88.7%), which is the degree of association for two binary

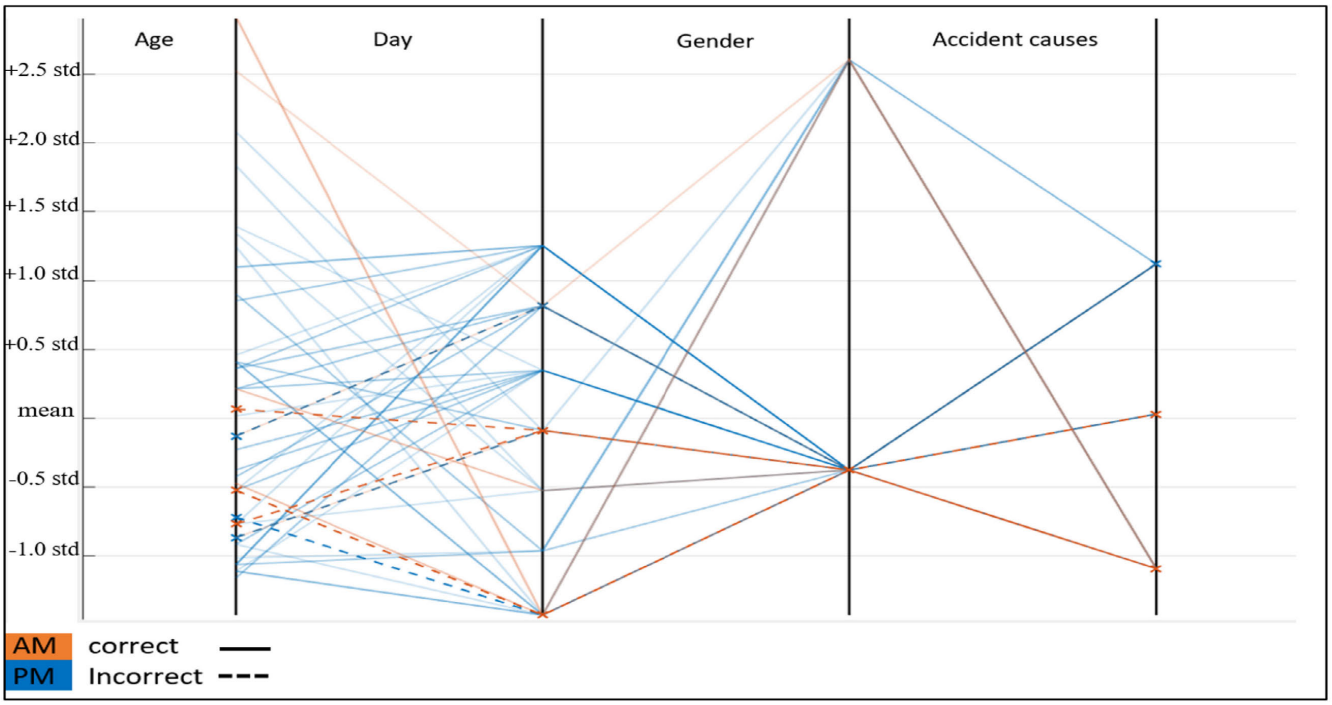

FIGURE 14. 2D parallel coordinates plot.

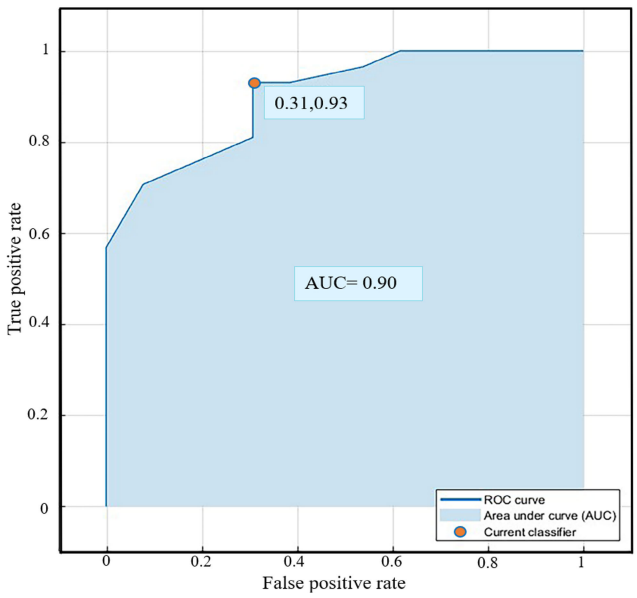

FIGURE 15. The ROC curve, which shows that the area under the curve (AUC) is 0.90. This evaluation model with positive class PM and a threshold of between 50% and 80% of the data (training data) vs. 20% of the data (testing dataset).

variables, are calculated utilising the MLT. The accuracy is the product of correct predictions over the total number of instances that have been evaluated. For a further investigation of features and visualisation of the prediction traits or correlation features of the results connecting to the accident patterns and safety predictors, see the parallel coordinate plot (Fig. 14).

In the model, the precision percentage and the recall rate indicate that the model had few false positives and negatives; hence, the model was more correct than incorrect when deciding whether the passengers involved in the accidents were there during the AM vs. PM. The area under the curve (AUC) was measured under the ROC curve. The decision tree achieves higher AUC values of 0.90, which indicates an improved classifier performance (Fig. 15).

VI. DISCUSSION

New technologies, such as ML, are utilized in numerous methods that can improve the safety of railways, manage risks in stations, and address accidents even outside of stations. For evolving and testing ML technology, a handful of accidents in railway stations are used, followed by training and testing datasets. Analysing the history of accidents can be performed locally or internationally and presents the root cause of the incidents and the correlations between many factors in different systems. From the model and case study, it is clear that applying ML modelling to railway station safety is a challenge, and more in-depth technical and analytical work is needed; therefore, more in-depth research is required. The accessibility to the details of the accidents presents challenges, such as privacy and availability challenges, for processing safety data in real time. We must integrate some of the systems in the railway stations and possibly automate data gathering to extract the most useful data. The railway industry can choose any safety dataset that has been recorded to teach an ML application with a range of analytical methods. Additionally, they can select safety datasets for analysis and

validate other such user behaviours or ticketing systems to determine any correlation and thus design predictors. It has been noted that the platform is a significant area of the station where many accidents occur, and the train interfacing with the passengers is a key aspect of the selected accidents. Some factors, such as time, where PM saw more accidents than AM, are clear. A high ratio of accidents occurring for males has been seen in two important age groups: the young and the old. Many factors, such as intoxication, must be considered; however, in the research, we attempt to apply ML rather than any deep analysis of the causes of the accidents. Several factors need to be involved in understanding the entire image of the accidents that have not been available in many open-source datasets. The DT method has been applied, and it shows how the prediction path predicts the target instances. The results present acceptable values that will be further justified with more available data as a part of future research. Depending on the selected data type and the prediction targets, the proper algorithms can be chosen. The classification of supervised ML has been applied and presented, therein showing high performance; some of the objectives of the model are as follows:

1 - Providing information that may demand that future railway stations perform in-depth analysis and classification and consider how they can obtain automated safety, therein being integrated with other developments or advanced techniques.

2 - Determining any possible shorting in current safety systems or frameworks and then improving the comprehensiveness of any sophisticated technology.

3 - Prediction of risk or consequences based on official recorded safety data.

The methodology of Ml is a promising technique that can learn from historical data and overcome uncertainty. In addition, the method affords real-time output to the decision maker and opens new windows to the cloud, IoT, smart stations and smart cities. This method can be used in real time to present the situation in the station in a timely manner. The technique leads to automation of the field and allows the process to be smarter. The ML technique can be fed with data by integrating many systems, such as automated fare collection (AFC) systems, fire and alarm systems, and any external systems such as police and other agencies, as well as safety record systems from other stations [30], [69]. Finally, the intelligent analytical approach used in this research yields more beneficial knowledge of rail station safety and will be useful in the future for designing risk management plans for rail stations worldwide.

VII. CONCLUSION

Various ML methods can be applied to safety tasks in the railway industry. In this study, an innovative proposal to utilise the true potentials of ML by the railway industry for improving the safety of stations is presented. Based on the study in this paper, the supervised algorithm performs accurately, and state-of-the-art applications can be effectively addressed using ML. Additionally, employing a variety of algorithms using ML provides robust and beneficial analysis of the history of the safety records. We have demonstrated the applicability of DTs to this safety task for railway stations. Although there are other classifiers with conceivably beneficial classifications and prediction performances, DTs yield easily interpretable accident details. The MLT demonstrated the validity of the model and the distributed analysis of the data. Additionally, it was employed to determine the relevance and importance of the chosen accident conditions. This method achieves good prediction accuracy, in this case, and we used a rather small dataset to prove the application of ML in railway station safety, where there is no doubt that larger datasets and more attributes would play a significant role in the analysis and results. The classification of supervised ML has been applied and presented in this study, therein showing high and acceptable performance. Indeed, a practical application requires a huge amount of test data and accident details for teaching the model and thus producing more patterns and predictions. From the model and case study, applying ML modelling for improving safety in railway stations is a challenge, and deeper technical and analytical work is needed. Therefore, more in-depth research is required. To be able to process safety data in real time, we have to integrate some of the systems in the railway stations and possibly use automated gathering of the data to extract the most useful data from many indicators. The railway industry can choose any safety data sets that have been recorded to teach an ML application with a range of analytical methods. Additionally, they can select safety datasets for analysis and validate other aspects, such as user behaviours and ticket systems, to determine any correlation and to design predictors. It has been noted that the platform is a significant area of the station where many accidents occur, and the train interfacing with the passengers is the key to the selected accidents. Finally, predicting people's behaviours and accident conditions is strategically of great value in safety and security. In general, but also specifically in the railway industry, this topic may be addressed by ML in the near future. However, the shortage of data available to apply ML remains a challenge for researchers. Moreover, in this work, accidents were not only analysed, but also a method was recommended to enhance ML applications for railway safety, risk management and accident investigation conceptualization, implementation, and big data. We hope that such proposals will greatly benefit future research concerning ML in railway safety research.

REFERENCES

[1] H. J. Parkinson, G. Bamford, and B. Kandola, "The development of an enhanced bowtie railway safety assessment tool using a big data analytics approach," in *Proc. Int. Conf. Railway Eng. (ICRE)*, Brussels, Belgium, May 2016, pp. 1–9, doi: 10.1049/cp.2016.0510.

[2] C. Martin and H. Leurent, *Technology and Innovation for the Future of Production: Accelerating Value Creation*. Geneva, Switzerland: World Economic Forum, 2017.

[3] R. Xu, J. Han, W. Qi, J. Meng, and H. Zhang, "Railway fastener image recognition method based on multi feature fusion," *IOP Conf. Mater. Sci. Eng.*, vol. 397, Aug. 2018, Art. no. 012119, doi: 10.1088/1757-899x/397/1/012119.

[4] L. Dai, "A machine learning approach for optimisation in railway planning," Ph.D. dissertation, Dept. Math. Comput. Sci., Delft Univ. Technol., Delft, The Netherlands, 2018.

[5] C. Zuo, Y. Zhang, Z. Xing, and Y. Qin, "Study on safety evaluation of urban rail transit station," in *Proc. 33rd Chin. Control Conf.*, Nanjing, China, Jul. 2014, pp. 3187–3190, doi: 10.1109/ChiCC.2014.6895462.

[6] (Apr. 2018). *Network Rail, Digital Railway Strategy, report*. Accessed: Nov. 18, 2019. [Online]. Available: https://cdn.networkrail.co.uk/wp-content/uploads/2018/05/Digital-Railway-Strategy.pdf

[7] (2018). *Network Rail, The Digital Railway Programme*. Accessed: Nov. 18, 2019. [Online]. Available: https://cdn.networkrail.co.uk/wp-content/uploads/2018/05/Digital-Railway-Programme.pdf

[8] J. E. Diekmann, "Risk analysis: Lessons from artificial intelligence," *Int. J. Project Manage.*, vol. 10, pp. 75–80, May 1992, doi: 10.1016/0263-7863(92)90059-I.

[9] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[10] N. Paltrinieri, L. Comfort, and G. Reniers, "Learning about risk: Machine learning for risk assessment," *Saf. Sci.*, vol. 118, pp. 475–486, Oct. 2019, doi: 10.1016/j.ssci.2019.06.001.

[11] C. Bergmeir, G. Sáinz, C. M. Bertrand, and J. M. Benítez, "A study on the use of machine learning methods for incidence prediction in high-speed train tracks," in *Recent Trends in Applied Artificial Intelligence* (Lecture Notes in Computer Science). Berlin, Germany: Springer, Jun. 2013, pp. 674–683, doi: 10.1007/978-3-642-38577-3_70.

[12] A. Thaduri, D. Galar, and U. Kumar, "Railway assets: A potential domain for big data analytics," *Procedia Comput. Sci.*, vol. 53, pp. 457–467, Jan. 2015, doi: 10.1016/j.procs.2015.07.323.

[13] H. Li, "Improving rail network velocity: A machine learning approach to predictive maintenance," *Transp. Res. C, Emerg. Technol.*, vol. 45, pp. 17–26, Aug. 2014, doi: 10.1016/j.trc.2014.04.013.

[14] S. Yella, S. Ghiamati, and M. Dougherty, "Condition monitoring of wooden railway sleepers using time-frequency techniques and pattern classification," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, San Antonio, TX, USA, Oct. 2009, pp. 4164–4169, doi: 10.1109/ICSMC.2009.5346713.

[15] R. G. Nyberg, "A machine learning approach for recognising woody plants on railway trackbeds," in *Proc. Int. Conf. Railway Eng. (ICRE)*, Brussels, Belgium, 2016, pp. 1–5, doi: 10.1049/cp.2016.0513.

[16] H. Mukojima, "Moving camera background-subtraction for obstacle detection on railway tracks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 3967–3971, doi: 10.1109/ICIP.2016.7533104.

[17] X. Gibert, V. M. Patel, and R. Chellappa, "Deep multitask learning for railway track inspection," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 153–164, Jan. 2017, doi: 10.1109/TITS.2016.2568758.

[18] L. Hajibabai, "Wayside defect detector data mining to predict potential WILD train stops," in *Proc. Annu. Conf. Expo. Amer. Railway Eng. Maintenance-Way Assoc. (AREMA)*, Chicago, IL, USA, Sep. 2012, pp. 1–39.

[19] G. Krummenacher, C. S. Ong, S. Koller, S. Kobayashi, and J. M. Buhmann, "Wheel defect detection with machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1176–1187, Apr. 2018, doi: 10.1109/TITS.2017.2720721.

[20] A. Núñez, J. Hendriks, Z. Li, B. D. Schutter, and R. Dollevoet, "Facilitating maintenance decisions on the Dutch railways using big data: The ABA case study," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Washington, DC, USA, Oct. 2014, pp. 48–53, doi: 10.1109/BigData.2014.7004431.

[21] A. Jamshidi, S. Faghih-Roohi, S. Hajizadeh, A. Núñez, R. Babuska, R. Dollevoet, Z. Li, and B. De Schutter, "A big data analysis approach for rail failure risk assessment," *Risk Anal.*, vol. 37, no. 8, pp. 1495–1507, Aug. 2017, doi: 10.1111/risa.12836.

[22] X. Zhang and D. Gong, "Application of big data technology in marketing decisions for railway freight," in *Proc. Int. Conf. Logistics Eng. Manage. (ICLEM)*, Shanghai, China, Sep. 2014, pp. 1136–1141, doi: 10.1061/9780784413753.172.

[23] F. Ghofrani, Q. He, R. M. P. Goverde, and X. Liu, "Recent applications of big data analytics in railway transportation systems: A survey," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 226–246, May 2018, doi: 10.1016/j.trc.2018.03.010.

[24] E. N. Martey, L. Ahmed, and N. Attoh-Okine, "Track geometry big data analysis: A machine learning approach," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Boston, MA, USA, Dec. 2017, pp. 3800–3809, doi: 10.1109/BigData.2017.8258381.

[25] X. Chen, "Railway passenger volume forecasting based on support vector machine and genetic algorithm," in *Proc. ETP Int. Conf. Future Comput. Commun.*, Wuhan, China, Jun. 2009, pp. 282–284, doi: 10.1109/FCC.2009.81.

[26] J. Peters, B. Emig, M. Jung, and S. Schmidt, "Prediction of delays in public transportation using neural networks," in *Proc. Int. Conf. Comput. Intell. Model. Control Automat. Int. Conf. Intell. Agents Web Technol. Internet Commerce (CIMCA-IAWTIC)*, Vienna, Austria, Nov. 2005, pp. 92–97, doi: 10.1109/CIMCA.2005.1631451.

[27] L. Bai, R. Liu, Q. Sun, F. Wang, and F. Wang, "Classification-learning-based framework for predicting railway track irregularities," *Proc. Inst. Mech. Eng. F, J. Rail Rapid Transit*, vol. 230, no. 2, pp. 598–610, Oct. 2014, doi: 10.1177/0954409714552818.

[28] J. M. Hart, L. F. Molina, E. Resendiz, J. R. Edwards, N. Ahuja, and C. P. L. Barkan, "Development of a machine vision system for the inspection of heavy-haul railway turnout and track components," in *Proc. Int. Heavy Haul Assoc. Conf.*, Calgary, AB, Canada, Jul. 2011, pp. 1–8.

[29] C. Hu and X. Liu, "Modeling track geometry degradation using support vector machine technique," in *Proc. Joint Rail Conf. Amer. Soc. Mech. Eng. Digit. Collection*, Columbia, SC, USA, Apr. 2016, doi: 10.1115/JRC2016-5739.

[30] H. Alawad and S. Kaewunruen, "Wireless sensor networks: Toward smarter railway stations," *Infrastructures*, vol. 3, no. 3, p. 24, Jul. 2018, doi: 10.3390/infrastructures3030024.

[31] S.-C. Oh, G.-D. Kim, W.-T. Jeong, and Y.-T. Park, "Vision-based object detection for passenger's safety in railway platform," in *Proc. Int. Conf. Control Automat. Syst.*, Seoul, South Korea, Oct. 2008, pp. 2134–2137, doi: 10.1109/ICCAS.2008.4694449.

[32] E. Bikov, P. Boyko, E. Sokolov, and D. Yarotsky, "Railway incident ranking with machine learning," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Cancun, Mexico, Dec. 2017, pp. 601–606, doi: 10.1109/ICMLA.2017.00-95.

[33] A. Jamshidi, "A decision support approach for condition-based maintenance of rails based on big data analysis," *Transp. Res. C, Emerg. Technol.*, vol. 95, pp. 185–206, Oct. 2018, doi: 10.1016/j.trc.2018.07.007.

[34] Y. Han, Z. Liu, Y. Lyu, K. Liu, C. Li, and W. Zhang, "Deep learning-based visual ensemble method for high-speed railway catenary clevis fracture detection," *Neurocomputing*, Apr. 2019, doi: 10.1016/j.neucom.2018.10.107.

[35] H. Tam, K. Lee, S. Liu, L. Cho, and K. Cheng, "Intelligent optical fibre sensing networks facilitate shift to predictive maintenance in railway systems," in *Proc. Int. Conf. Intell. Rail Transp. (ICIRT)*, Singapore, Dec. 2018, pp. 1–4, doi: 10.1109/ICIRT.2018.8641602.

[36] M. Heidarysafa, K. Kowsari, L. Barnes, and D. Brown, "Analysis of railway accidents' narratives using deep learning," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Orlando, FL, USA, Dec. 2018, pp. 1446–1453, doi: 10.1109/ICMLA.2018.00235.

[37] H. Rowshandel, G. L. Nicholson, J. L. Shen, and C. L. Davis, "Characterisation of clustered cracks using an ACFM sensor and application of an artificial neural network," *NDT E Int.*, vol. 98, pp. 80–88, Sep. 2018, doi: 10.1016/j.ndteint.2018.04.007.

[38] Y. Jiang, H. Wang, G. Tian, Q. Yi, J. Zhao, and K. Zhen, "Fast classification for rail defect depths using a hybrid intelligent method," *Optik*, vol. 180, pp. 455–468, Feb. 2019, doi: 10.1016/j.ijleo.2018.11.053.

[39] H. Wang, A. Núñez, Z. Liu, D. Zhang, and R. Dollevoet, "A Bayesian network approach for condition monitoring of high-speed railway catenaries," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2019.2934346.

[40] The BigML Team. Corvallis, OR, USA. (2019). *Classification and Regression With The BigML Dashboard*. Accessed: Nov. 18, 2019. [Online]. Available: http://bigml.com

[41] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," in *Proc. Conf. Emerg. Artif. Intell. Appl. Comput. Eng. Real Word AI Syst. Appl. eHealth HCI Inf. Retr. Pervasive Technol.*, vol. 160, 2007, pp. 3–24, doi: 10.1007/s10462-007-9052-3.

[42] S. K. Chandrinos, G. Sakkas, and N. D. Lagaros, "AIRMS: A risk management tool using machine learning," *Expert Syst. Appl.*, vol. 105, pp. 34–48, Sep. 2018, doi: 10.1016/j.eswa.2018.03.044.

[43] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[44] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, "Analysis of various decision tree algorithms for classification in data mining," *Int. J. Comput. Appl.*, vol. 163, no. 8, pp. 15–19, Apr. 2017, doi: 10.5120/ijca2017913660.

[45] L. Breiman, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth International Group, 1984.
[46] L. Schnieder, E. Schnieder, and T. Ständer, "Railway safety and security—Two sides of the same coin?!" in *Proc. Int. Railway Saf. Conf.*, Braunschweig, Germany, 2009, pp. 1–16.
[47] V. Villa, N. Paltrinieri, and V. Cozzani, "Overview on dynamic approaches to risk management in process facilities," *Chem. Eng. Trans.*, vol. 43, pp. 2497–2502, Jan. 2015.
[48] A. Meel and W. D. Seider, "Plant-specific dynamic failure assessment using Bayesian theory," *Chem. Eng. Sci.*, vol. 61, no. 21, pp. 7036–7056, Nov. 2006, doi: 10.1016/j.ces.2006.07.007.
[49] G. D. Creedy, "Quantitative risk assessment: How realistic are those frequency assumptions?" *J. Loss Prevention Process Ind.*, vol. 24, no. 3, pp. 203–207, May 2011, doi: 10.1016/j.jlp.2010.08.013.
[50] S. Andersen and B. A. Mostue, "Risk analysis and risk management approaches applied to the petroleum industry and their applicability to IO concepts," *Saf. Sci.*, vol. 50, no. 10, pp. 2010–2019, Dec. 2012, doi: 10.1016/j.ssci.2011.07.016.
[51] M. Kalantarnia, F. Khan, and K. Hawboldt, "Dynamic risk assessment using failure assessment and Bayesian theory," *J. Loss Prevention Process Ind.*, vol. 22, no. 5, pp. 600–606, Sep. 2009, doi: 10.1016/j.jlp.2009.04.006.
[52] X. Yang, S. Haugen, and N. Paltrinieri, "Clarifying the concept of operational risk assessment in the oil and gas industry," *Saf. Sci.*, vol. 108, pp. 259–268, Oct. 2018, doi: 10.1016/j.ssci.2017.12.019.
[53] C. Q. X. Poh, C. U. Ubeynarayana, and Y. M. Goh, "Safety leading indicators for construction sites: A machine learning approach," *Automat. Construct.*, vol. 93, pp. 375–386, Sep. 2018, doi: 10.1016/j.autcon.2018.03.022.
[54] (Feb. 2, 2017). *How Big Data, Modelling and Machine Learning are Shaping Safety in BC*. Accessed: Nov. 18, 2019. [Online]. Available: https://www.technicalsafetybc.ca/blog/how-big-data-modeling-and-machine-learning-are-shaping-safety-bc
[55] V. Vapnik, "Principles of risk minimization for learning theory," in *Proc. 4th Int. Conf. Neural Inf. Process. Syst.*, Denver, CO, USA, Dec. 1991, pp. 831–838.
[56] K. R. Varshney, "Engineering safety in machine learning," in *Proc. Inf. Theory Appl. Workshop (ITA)*, La Jolla, CA, USA, Jan. 2016, pp. 1–5, doi: 10.1109/ITA.2016.7888195.
[57] G. Cai, L. Jia, L. Zhou, Y. Liang, and X. Li, "Research on rail safety security system," *Int. J. Econ. Manage. Eng.*, vol. 4, no. 8, pp. 1938–1943, 2010. Accessed: Nov. 18, 2019. [Online]. Available: https://publications.waset.org
[58] An Agency of the European Union, *Big Data In Railways*, document ERA-PRG-004-TD-003 V 1.0/, European Union Agency Railways, Valenciennes, France, 2016. Accessed: Nov. 18, 2019. [Online]. Available:https://www.era.europa.eu/
[59] Government Agency. Derby, Farnborough, U.K. *Rail Accident Investigation Brach (RAIB)*. Accessed: Nov. 18, 2019. [Online]. Available: https://www.gov.uk/government/organisations/rail-accident-investigation-branch
[60] S. Carpenter, *Passenger Risk at the Platform-Train Interface*. London, U.K.: Rail Safety and Standards Board, 2011.
[61] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Appl. Artif. Intell.*, vol. 17, nos. 5–6, pp. 375–381, May 2003, doi: 10.1080/713827180.
[62] R. Houari, A. Bounceur, M. T. Kechadi, A. K. Tari, and R. Euler, "Dimensionality reduction in data mining: A copula approach," *Expert Syst. Appl.*, vol. 64, pp. 247–260, Dec. 2016, doi: 10.1016/j.eswa.2016.07.041.
[63] M. S. Hajakbari and B. Minaei-Bidgoli, "A new scoring system for assessing the risk of occupational accidents: A case study using data mining techniques with Iran's Ministry of Labor data," *J. Loss Prevention Process Ind.*, vol. 32, pp. 443–453, Nov. 2014, doi: 10.1016/j.jlp.2014.10.013.
[64] N. Leavitt, "Bringing big analytics to the masses," *Computer*, vol. 46, no. 1, pp. 20–23, Jan. 2013. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/MC.2013.9, doi: 10.1109/mc.2013.9.
[65] H. Liu, Z. Liu, and D. Liu, "Application of machine learning methods in maritime safety information classification," in *Proc. 10th Int. Conf. Adv. Comput. Intell. (ICACI)*, Xiamen, China, Mar. 2018, pp. 735–740, doi: 10.1109/ICACI.2018.8377552.
[66] M. Chen, A. X. Zheng, J. Lloyd, M. I. Jordan, and E. Brewer, "Failure diagnosis using decision trees," in *Proc. Int. Conf. Autonomic Comput.*, New York, NY, USA, May 2004, pp. 36–43, doi: 10.1109/ICAC.2004.1301345.
[67] Y. Yuan and M. J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets Syst.*, vol. 69, no. 2, pp. 125–139, Jan. 1995, doi: 10.1016/0165-0114(94)00229-Z.
[68] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Denver, CO, USA, Oct. 2015, pp. 1322–1333, doi: 10.1145/2810103.2813677.
[69] H. Alawad, S. Kaewunruen, and A. Min, "Utilizing big data for enhancing passenger safety in railway stations," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 603, Sep. 2019, Art. no. 052031, doi: 10.1088/1757-899X/603/5/052031.
[70] S. Kaewunruen, J. M. Sussman, and A. Matsumoto, "Grand challenges in transportation and transit systems," *Frontiers Built Environ.*, vol. 2, p. 4, Feb. 2016, doi: 10.3389/fbuil.2016.00004.

HAMAD ALAWAD received the bachelor's degree in industrial engineering from King Saud University and the master's degree in fire and safety engineering, U.K. He is currently pursuing the Ph.D. degree with the Birmingham Centre for Railway Research and Education, University of Birmingham, U.K. His research interests include safety, machine learning, and artificial intelligence.

SAKDIRAT KAEWUNRUEN received the Ph.D. degree in structural engineering from the University of Wollongong, Australia. He has expertise in transport infrastructure engineering and management, and successfully dealing with all stages of infrastructure life cycle and assuring safety, reliability, resilience, and sustainability of rail infrastructure systems. He is a chartered engineer. He has over 400 technical publications.

MIN AN is a Professor of construction and risk management with the University of Salford. He is also a Visiting Professor with Beijing Jiaotong University, China. His research work has been funded from a variety of sources including EPSRC, EU, government agencies, and industry. He is/was the PI of 25 financed contracts.