applied sciences

MDPI

Article

Dynamic Railway Derailment Risk Analysis with Text-Data-Based Bayesian Network

Liu Yang [1], Keping Li [1,*], Guozheng Song [2] and Faisal Khan [3]

[1] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China; lyang90@bjtu.edu.cn
[2] Functional Safety Center, Instrumentation Technology and Economy Institute, Beijing 100055, China; gs1870@mun.ca
[3] Centre for Risk, Integrity and Safety Engineering (C-RISE), Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1B 3X5, Canada; fikhan@mun.ca
* Correspondence: kpli@bjtu.edu.cn

Abstract: In recent years, transportation system safety analysis has become increasingly challenging and highly demanding. Unstructured data contain sufficient information from which inherent interactions can be extracted. Determining how to process and fuse a large amount of unstructured data is a challenging task. In this paper, we propose a text-based Bayesian network (TBN) method to establish a Bayesian network (BN) based on text records, where the BN's arcs are obtained from barrier relationships identified by a graphical model and its prior probabilities stem from fault trees. The comparative experimental results illustrate that the text-based method in TBN is efficient. The precision, recall and F-measure of TBN are 8.64%, 10.70% and 9.84% higher, respectively, than the most frequent (MF) result. Moreover, compared to the traditional BN, whose prior probabilities are frequently acquired from experts, the prior probabilities of the proposed text-based BN (TBN) have a high confidence. The experimental results of a train derailment accident case study show that with changes in the train derailment probabilities and the safety potentials of the barriers, the TBN generates quantitative results and reveals the critical risks of derailment accidents. Additionally, this work demonstrates relevant nonlinear relationships to improve the assessment results. Therefore, based on text-based data, this study reveals that barrier safety analysis has the potential to identify high-risk barriers, which can guide managers to enhance these barriers.

Keywords: railway derailment; barrier safety; bayesian network; fault trees; text analysis

Citation: Yang, L.; Li, K.; Song, G.; Khan, F. Dynamic Railway Derailment Risk Analysis with Text-Data-Based Bayesian Network. *Appl. Sci.* **2021**, *11*, 994. https://doi.org/10.3390/app11030994

Received: 17 December 2020
Accepted: 19 January 2021
Published: 22 January 2021

1. Introduction

Railway transportation has exhibited an upward trend in terms of the daily demand, especially because of its considerably lower price than air transportation and higher speed than shipping transportation. As a result that railway accidents can result in extreme economic losses, even casualties, railway safety has received increasing attention in recent years. To avoid serious accidents, numerous studies have introduced various accident causation models to analyze safety assessments. For example, to analyze railway accidents, the Systems Theoretic Accident Model and Process (STAMP) method was proposed to examine the accident spreading process founded on basic systems theory concepts [1] and to provide certain measures. Compared to event chain models [2], STAMP regards systems as interrelated components that are maintained in a state of dynamic equilibrium by information and control feedback loops. The Human Factors Analysis and Classification System (HFACS) based on the Swiss model was suggested to evaluate human factors (HFs) [3]. However, in this type of model, the system is considered as a whole by dividing the whole system into several subsystems or hierarchical levels, which explains accidents as interconnected events; however, their analysis results are insufficient to establish effective

cause control [4]. Moreover, systematic accident model-based methods are qualitative methods, not quantitative.

To overcome the drawbacks of systematic models, the theory of complex networks has been commonly applied in transportation. In a complex network, an accident causation network is necessarily established. Causation is represented by nodes, and the relationship between two nodes is shown as a link. By applying statistical characteristics, various models have been introduced by the node degree. Ma et al. established a network to analyze or assess the risks in railway systems [5]. Due to the lack of weighting in much of the work based on complex networks, Zhou et al. improved Ma's model. Zhou et al. proposed a complex network-based method that considers the weights of edges [6]. However, these studies have not considered cascading processes. To overcome this limitation, Zhou et al. proposed a railway fault spreading model based on the dynamics of the process of failure interactions in networks [7]. Li and Wang developed a model based on a complex network for risk monitoring [4]. This model has been mainly employed to identify accident causal factors and to analyze how these factors influence each other. Liu and Li [8] recommended a cascading failure model considering cascading processes.

Many other quantitative models based on fault trees have also been introduced to analyze safety assessments. As a result that it is difficult to capture the failure probabilities of basic events based on past experience using fault trees, the fuzzy set theory has been adopted to avoid any potential shortcomings. Liu et al. established the fault tree analysis method combined with quantitative analysis, in which the failure probabilities of basic events are assumed to be intuitionistic trapezoidal fuzzy numbers due to the incompleteness of the prior information and the complexity of the decision environment [9]. Dindar et al. proposed a method for the quantification and evaluation of fault trees in fuzzy environments [10]. It was designed for rare events and the identification of the probability and the underlying root cause of derailment. Ref. [11] proposed a method to calculate the probability of adverse events using historical information about transport cases for the learning process of Artificial Neural Networks (ANN) and Machine Learning (ML), which can be used to develop a system for identifying high risk places and potential risks.

Bayesian networks (BNs) possess dynamic properties and nonlinear characteristics and can demonstrate dependency. BNs have been widely applied to investigate accidents in recent years. Song et al. established a bow-tie model to illustrate the occurrence and escalation process of occupational accidents, slips, trips and falls from heights (STFs), and the results indicate the BNs' advantage in terms of quantifying occupational risks [12]. Reference [13] suggested the subsequent mapping of fault trees into BNs to overcome the limited predictive capacity of fault trees. Chen et al. showed that a BN could correctly reason through door system faults, and the results provide a reference and some advice for fault diagnosis and maintenance of railway vehicle doors [14]. To better apply the BN, a large number of models have been developed in various fields, such as the process, transportation, oil and ocean industries. Rathnayaka et al. [15,16] proposed the System Hazard Identification, Prediction and Prevention (SHIPP) methodology, which is a systematic and comprehensive safety analysis procedure. SHIPP describes a developed process accident model with a predictive capacity. It uses a combination of event and fault tree concepts to model the cause-consequence relationship and captures the process operational behavior and updates the accident likelihood using the Bayesian updating mechanism. The application of Bayesian network has been widely used in many random events determined by diffuse factors, for example, transportation process is performed between the initial nodes (depots) and end nodes (tram reversing loops) in [17]. To analyze the nonlinear interactions between accident contributory factors, Dindar et al. [18] presented a new nonsequential barrier-based process accident model. The conditional dependencies among accident contributory factors within prevention barriers are examined with the BN under various relaxation strategies and nonsequential failure prevention (safety) barriers. A BN combines the two theoretical systems of the probability and graph theories and exhibits an excellent inference capacity with uncertainty and knowledge representation [3].

Song et al. [19] developed a BN model with dynamic features and dependency among barriers using nonlinear and nonsequential accident models for process systems. Motivated by these studies, the BN model will be adopted in this paper due to its dynamic properties and nonlinear characteristics.

Keywords thought that it can convey important information within a few words. In addition, keywords can reflect the key points of an article or the author's writing intention. With the rapid development of Natural Language Process(NLP), a growing number of fundamental methods on extracting or depicting text keywords are proposed. Term frequency-inverse document frequency (TF-IDF) was proposed by Salton and Buckley in 1988. It contains two statistics terms frequency(TF) and inverse document frequency (IDF), of which TF is applied to measure the popularity and IDF is applied to measure the discrimination between the target files and the background files [20]. In 2012, most frequent (MF) was proposed to count document frequency for all of the words that appearing within the document. TF-IDF and MF are often used as benchmarks [21]. To prevent words with a very low frequency from getting a high score by chance, Kuhn et al. extended a frequency method in citation network as its sticking factor divided by its vital factor [22]. As network-based method showed the best performance in extracting keywords [23], and Yang et al. proposed a new network model based on a complex network. The method considers the influence of sentences and the relationship between a word and sentence [24]. In the field of transportation, Yang et al. proposed a method to identify the cause factors from railway text reports in 2019 [25]. However, Ref. [25] did not discuss how to apply the cause factors where are from text data to risk analysis in the transportation field. Thus, in this paper, the most important task was to explore the application of the keywords extraction method in risk analysis.

Specifically, we first define the system and the barriers to establish a graphical model. The graphical model not only reveals the accident occurrence principle and process but also shows the relations between barriers, which are the fundamental relations in BNs. Second, we perform text processing and characterize the text features and then establish fault trees based on the text representation information. In this step, the prior probabilities are determined. Next, based on the graphical model, we establish a BN (the text-based BN or TBN) using barriers, in which the relations are retrieved from the graphical model and the prior probabilities are obtained from the fault trees. Then, we perform a quantitative safety assessment of railway systems using the BN. This approach could help identify the key factors of accidents. Additionally, a dynamic BN shows the status changes of the barriers when certain conditions change. The method establishes the dependency of different parts, and, therefore, it could update the state of one subsystem using data from another subsystem. This ensures that managers obtain the latest state information to the greatest extent, thus facilitating risk reduction. The work demonstrates the inherent nonlinear relationships to improve the assessment results. We also study the barrier safety potential to determine high-risk barriers, which could provide information to managers about which barriers should be prioritized.

The main contributions of our work are as follows:

- to reduce the data uncertainty due to the prior probabilities of the BN by establishing fault trees to obtain relevant prior probabilities from numerous text data. In contrast, prior probabilities have been obtained in previous work from expert experience in practical applications, which increases the uncertainty;
- to quantitatively and accurately evaluate railway safety. The analysis results not only represent the quantitative risk but also reveal the key barriers and critical potential risks of defects;
- the proposed method has the advantage that the determined cause chains considering safety barriers better fit the accident characteristics than the cause chains directly determined from text records.

This paper is organized as follows: Section 2 introduces fundamental information and describes the methodology. Section 3 presents the detailed steps of the TBN method.

In Section 4, a case study is examined, and a corresponding graphical model, fault trees and BN are then established. Section 5 provides conclusions.

2. Methodology

First, fundamental information must be provided. The BN is used because it has dynamic properties and nonlinear characteristics and uncovers dependencies. Text data are adopted because they provide objective prior data for the assessment model. A graphical model should be employed because it shows not only the relationships between barriers but also the accident occurrence principle of the process. Fault trees are established because they provide the prior probabilities of BNs from text records.

2.1. Text Extraction

The network-word-sentence (NWS) method is a keyword extraction method proposed in a previously published paper, which is based on the complex network theory [24]. The NWS method includes the following two layers: a sentence network in the upper layer and a word network in the lower layer. In the sentence network, a square represents a node (sentence), while in the word network, a circle represents a node (word). The relationship between similar nodes is represented by a solid line. When a node (word) in the word network coincides with a node (sentence) in the sentence network, their relationship can be built and represented by dotted lines. Therefore, we obtain the synthetic eigenvalue of each node (word) via the NWS method as follows:

$$E_i = \lambda WE_i + \gamma SWE_J \tag{1}$$

$$WE_i = \alpha \cdot bc_i + \beta \cdot cc_i \tag{2}$$

$$SWE_J = \omega_j \tag{3}$$

where E_i is the eigenvalue of word i, WE_i is the synthetic eigenvalue of word i, SWE_j is the eigenvalue of sentence j, c_i is the betweenness centrality, cc_i is the betweenness centrality, ω_j is the word node contribution from the sentence node contribution, $\lambda + \gamma = 1$ and $\alpha + \beta = 1 = 1$ [24].

2.2. Fault Tree

A fault tree, which is constructed from events and logic gates, is a graphical method to determine the failure probability of a complex system. The top event in the fault tree represents the major accident initiating the hazard, and the logic gates in the fault tree represent the numerous ways through which machine and human errors interact to cause the accident. In the AND gate, the failure probability of the top event is calculated with Equation (4), while in the OR gate, it is calculated with Equation (5), as follows:

$$p = \prod_{i=0}^{n} p_i \tag{4}$$

$$p = \prod_{i=0}^{n} (1 - p_i) \tag{5}$$

2.3. Bayesian Network (BN)

The BN represents the joint probability distribution P(U) of a set of discrete random variables U, incorporated in the network according to Equation (6) [25], as follows:

$$P(U) = \prod_{i=1}^{n} P(A_i | Pa(A_i)) \tag{6}$$

where $Pa(A_i)$ is the parent of variable A_i and $P(U)$ is the joint probability distribution of the variables.

3. Proposed Method: Text-Based Bayesian Network (TBN)

In this paper, a method with a combination of text records, a graphical model and a Bayesian network was proposed. Generally, the task of our proposed method contains three main phases, including graphical model establishment, fault trees establishment and Bayesian network establishment. To obtain a graphical model, the process of train moving and definitions of the safety barriers must be given at first. Then, the relationships between barriers from the graphical model can be used to establish fault trees from text-based data. After obtaining the relationships from the graphical model and the prior possibilities from the fault trees, the Bayesian network can be established. (see Figure 1). Since each record of derailment accidents is written by various persons working in different workplaces when an accident occurres, to help reduce ambiguity and to better facilitate a functional demonstration of the proposed method, some assumptions exist in this study, as follows.

1. The process of train operation is highly dependent on the conductor (driver).
2. This study focuses only on the process in the explication of records.
3. This study focuses on the most frequent accident type, i.e., derailment.
4. All data come from the equipment, and the accident-highway situation is not considered.

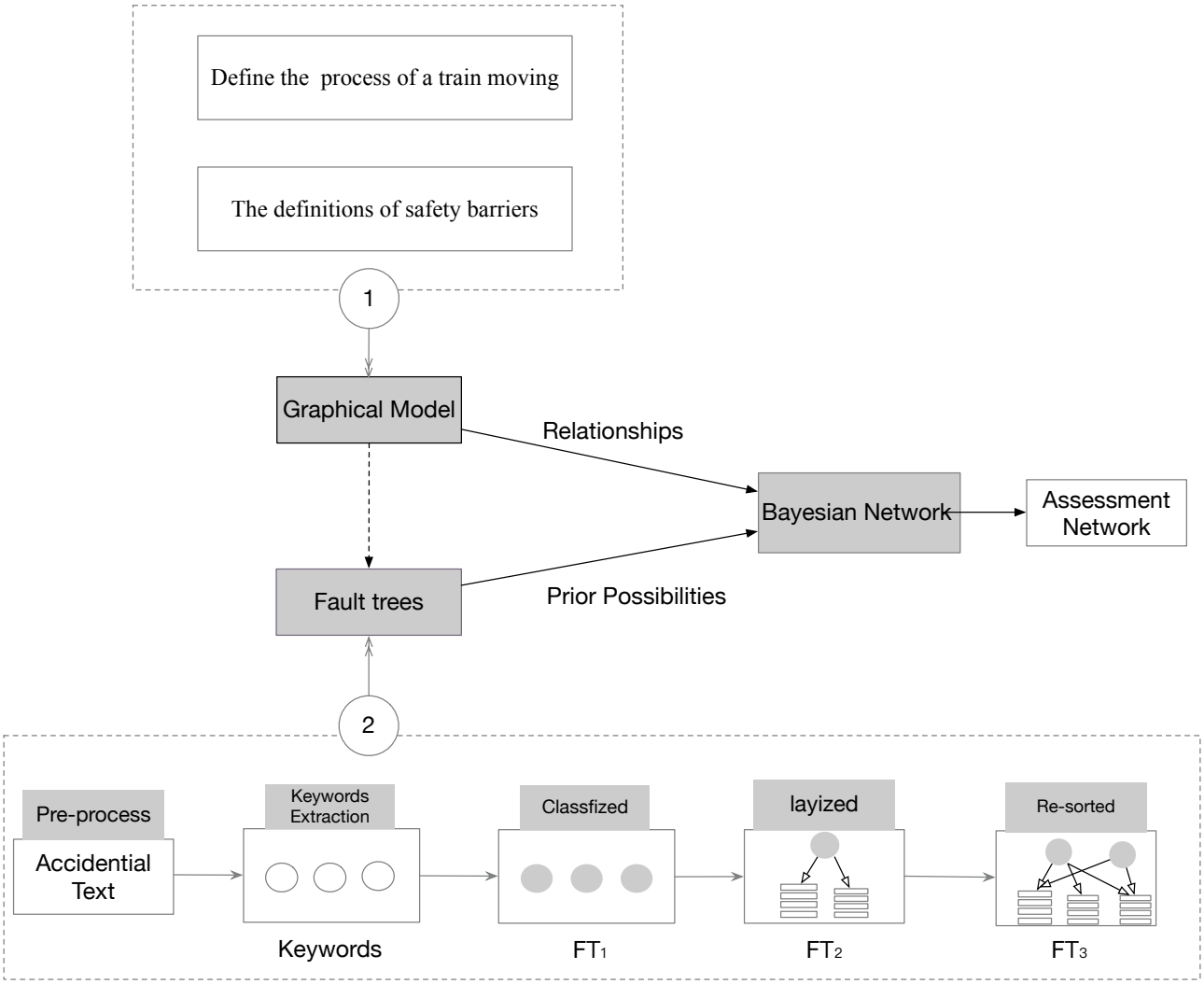


Figure 1. Framework of the proposed method.

3.1. Graphical Model Establishment

Although the Swiss model is applied to represent accident causation, it is able to represent only the linear relations of barriers. However, in reality, the barriers have nonlinear relationships, which means the barriers do not necessarily function one after another. To overcome this weakness, we proposed a graphical model representing the nonlinear relations between barriers. To establish a graphical model, the identification of barrier safety is the first urgent consideration.

First, we define the whole process of train derailment, i.e., from the travel plan to when the train stops after an accident occurs. To simplify the analysis, the process is divided into two stages, i.e., a preparation stage and an operating stage. Both stages contain several barriers to prevent a chain of failure from causing derailment. During the preparation stage, any train must undergo a series of inspections and maintenance operations by workers, drivers and so on before setting off. Thus, those inspections and maintenance operations could function as safety barriers to reduce the risk of derailment. Additionally, due to the rapid speed and high temperature that occur during the operation stage, some safety barriers play vital roles to decrease the growth of the possibility of derailment. Therefore, according to the process of a train from start to stop, six safety barriers are identified in an accident process: Initial Failure Prevention Barrier, Regular Inspection Barrier, Regular Maintenance Barrier, Process Failure Prevention Barrier, Failure Notice Barrier and Failure Control Barrier. The detailed definitions of the safety barriers and their practical meanings are as follows:

(1) Initial Failure Prevention Barrier

Initial failure means that the failure occurs in the train's preparation. In many records, if initial failures occur and are not to be observed, the failures could cause an accident to occur.

(2) Regular Inspection Barrier

Before leaving a station, a train must be inspected by workers to uncover any hidden failures. The inspections include daily inspections and regular inspections.

(3) Regular Maintenance Barrier

Once the train has been checked for failures, some measures must be taken to maintain the safe operation of the train.

(4) Process Failure Prevention Barrier

As the train starts, process failures occur while it is moving. In contrast to the initial failures, process failures come from external hazards instead of internal hazards, e.g., track damage, extreme weather and drunk drivers.

(5) Failure Notice Barrier

During the phase when the train is in motion, there are some obvious methods that can be used to find any hidden risks, e.g., the judgment of the driver, including his experience and eye vision, the use of a communication device to detect and announce risks or a plate that can show a failure, etc.

(6) Failure Control barrier

The function of this barrier is to prevent an accident when a failure is detected. If a barrier failure occurs, it means that the accident process is finished.

Table 1 provides the definitions of the barriers. As our study focuses on text data, to further the research, we first apply the text method and then establish fault trees based on the text records. Text features include the symbol features, semantic features, term length features and keyword characteristics discussed.

Table 1. Definitions of barriers.

Symbols	Meanings
B1	Initial Failure Prevention Barrier
B2	Regular Inspection Barrier
B3	Regular Maintenance Barrier
B4	Process Failure Prevention Barrier
B4a	Process Failure (Inner) Prevention Barrier
B4b	Process Failure (Outer) Prevention Barrier
B5	Failure Notice Barrier
B6	Failure Control Barrier

3.2. Classify, Layer and Re-Sort

The authors of [18] believed that failures could occur from a process such as the component-subsystem-system. A similar thought was also proposed in another work [4] (see Figure 2), considering the complexional train system, which obviously shows that a fault tree should be updated to close to the fault tree definition in the step. Therefore, we propose the next phrase progressively from FT1 to FT3, which is the final construction of the fault tree.

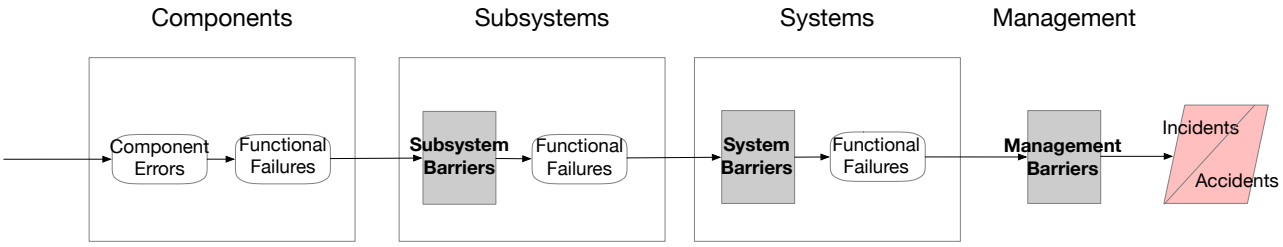


Figure 2. The barriers proposed in [4].

Next, the steps from the text record to the fault tree are as follows: Required: $CS = CI, CO, CH, CS$ represents the categories that the causes could belong to. CI is the internal factors, CO is the external factors and CH represents the human factors.

Step 1 Perform the text preprocessing and obtain the cut words.

In this step, each record contains its own cut words with a sequence of numbers, where set $R_i = \{cw_1, cw_2, ..., cw_j\}$, R_i is the set of cut words of the i-th record, and cw_j is the j-th cut word in the i-th record.

Step 2 Perform text feature analysis and feature extraction.

Based on the cause type classification, we divide these causes into the following three parts: external factors (O), internal factors (I) and human factors (H). From the perspective of train operation, the external factors are environmental factors, weather conditions and certain instruction signals. The HFs include the train driver, train passengers, administration staff and other people outside the train. Hence, we set the following:

$$\begin{aligned} IF &= \{ITF, IHF\} \\ OF &= \{EF, OTF, OHF\} \\ HF &= \{IHF, OHF\} \end{aligned} \tag{7}$$

where IF represents a set of internal factors, including train-factors (TF) and internal humans factors (IHF). OF represents a set of external factors, including environment factors, external track factors (OTF) and external humans factors (OHF). HF contains IHF and OHF.

Step 3 Obtain fault tree $FT(1)$.

First, we should determine the causes related to the HFs. If any word among the HFs occurs in a given record containing a cut word, we assign the record to group GHF.

Regarding to the records in group GHF, if any records primarily originate from train passengers, we place these records in group GIF, and they are defined as $GOF(OHF)$, i.e., the status of the driver and the reaction of the driver. Next, for the rest of the records, if EF weather words and OTF object words occur, we assign these records to $GOF(EF_i)$ and $GOF(OTF_i)$, respectively. Finally, the remaining records are assigned to group GIF. It is clear that the GIF group is not informative, so we must classify the contained records. Regarding the GIF group, if there is only one factor in IF, we define it as $GIF(IHF)$, but if both body factors and HFs occur, we define it as $GIF(ITF)$. Therefore, we obtain the following:

$$\begin{aligned} FT(1) &= \{GOF(1), GIF(1)\} \\ GOF(1) &= \{GOF(OTF), GOF(OHF)\} \\ GIF(1) &= \{GIF(ET), GIF(ITF), GIF(IHF)\} \end{aligned} \tag{8}$$

Step 4 Update and obtain $FT(2)$.

In this step, we must layer $FT(1)$ using the relation between word i and the state. In each subgroup of $GOF(1)$, we set word i as layer 1, and the following layer is a combination of all subgroups. For this combination, if word i and its previous or subsequent 2 words are the same, the subgroup is divided into several parts containing the same words. According to the assumptions above, all related words i in group $GOF(1)$ are defined as $GOF(2)$=$sub(w_i)$, $human(w_i)$, $other(w_i)$. Similar to GIF, we obey the assumptions and define $GIF(2)$=$sub(w_j)$, human (w_j) other (w_j). As such, we obtain $FT(2)$=$GOF(2)$, $GIF(2)$.

Step 5 Receive $FT(3) \leftarrow FT(2)$.

Following the definition principles of the system and barriers, there are four basic factors in each barrier, i.e., HFs, internal factors and external factors. For example, if barrier 1 contains the four HF conditions and external, internal and general factors, then we resort to fault tree FT(2) based on the barriers defined above. Hence, we obtain FT(3).

3.3. Establish the BN←FT(3)

In this step, a BN model will be established based on the above fault tree and safety barriers. Mapping fault trees into BNs is a simple process that is described below. With the conditional probability table (CPT) logic defined, the BN node probability is calculated through Equation (6), as follows:

$$P[X_1, X_2, ...X_n] = \prod_{i=1}^{n} P[X_i | Parent(X_i)] \tag{9}$$

As mentioned in Sections 3.1 and 3.2, the fault trees yield the prior probabilities of Barriers 1 and 3, and the graphical model illustrates the relationships between the barriers and certain conditional probabilities. In a BN, the nodes represent causal factors and target events, while the arcs reveal their dependence. The relationship between dependent nodes is represented by CPTs. In summary, the steps of our proposed model are as follows:

Step 1 Define the system and barriers and establish a graphical accident model. By analyzing the process of railway derailment, the train operation is divided into two stages. To reduce the probability of derailment, the safety barrier of each stage is given.

Step 2 Preprocess the accident text records.

Step 3 Obtain the fault trees from $FT(1)$ to $FT(2)$ based on the representation of the text records and barriers proposed in step 1.

Step 4 Determine the BN according to the barrier relationship given by the graphical model in step 1 and calculate the conditional probabilities of the prior probabilities and other barriers by the obtained $FT(3)$ in step 3.

Step 5 Conduct a quantitative safety assessment of the railway system using the acquired BN and identify the key accident factors. Moreover, the dynamic BN reveals the status changes of the barriers when a certain condition changes.

Step 6 Provide the cause chain of the accident.

4. Case Study and Experimental Results

In this study, the accident database is compiled based on 10-year railway equipment accident (REA) data from the Federal Railroad Administration (FRA) (http://www.fra.dot.gov), which include railway equipment and highway accident data, operational data and railroad casualty data. Train accidents are frequently the culmination of a sequence of events and a variety of conditions or circumstances that may have contributed to their occurrence, and they commonly meet certain reporting criteria. To better record and understand why an accident has occurred, a standard document named the Train Accident Cause Codes (TACC) is provided to describe the causal factors of the accident. In the TACC, there are 289 codes related to the causal factors of accidents. Figure 3 shows that the largest proportion is derailments, accounting for 59.32% of the 10-year records.

Hence, the main input of our proposed method are the text data of train derailment accidents, which are collected by FRA. Here, we selected the cause factors of derailment accidents as input. Five kinds of factors are defined as follows: (l) rack, roadbed and structures; (2) signal and communication; (3) mechanical and electrical failures; (4) train operation—human factors; (5) miscellaneous causes, including environment conditions, management, decisions and measures.

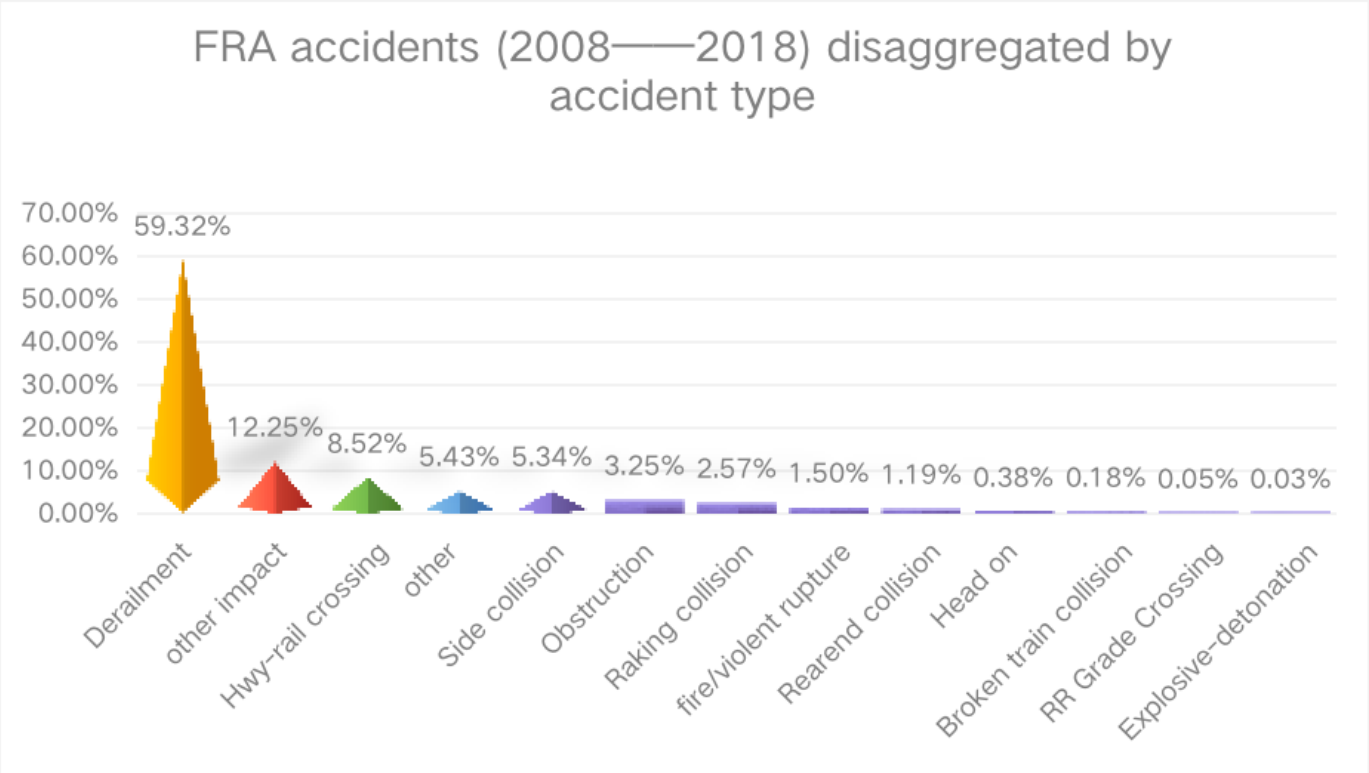


Figure 3. Federal Railroad Administration (FRA) accidents disaggregated by accident type.

To overcome the limitation of the BN model that it cannot straightforwardly identify related barriers, we use graphical models to help find out the barrier. In the graphical models, barriers have many contribution factors. To better understand the roles of barriers, we organized a structure to build fault trees (FTs) which are designed to identify the causal factors as in [13] where a suitable method was provided to map fault trees into BNs. Usually, the prior possibilities of the BN model are hard to decide, so we used a text based

method to obtain prior inputs from text data. Therefore, the great improvement of TBN is that a text-method was adopted to provide objective prior data for the assessment model.

First, we introduced an experiment to illustrate the efficiency of the text-based method in TBN. In this experiment, we selected network-word-sentence (NWS) and the most frequency (MF) as the text-based methods. NWS was proposed in our previous paper and is based on complex network theory, and the MF method has exhibited a prominent performance in the literature [24] and is frequently used. Then, we compared these two methods in all three metrics: precise, recall and F-measure. The performance of the NWS method has a precision of 58.77%, a recall of 61.86% and an F-measure of 60.28%. The precision, recall and F-measure of TBN are 8.64%, 10.70% and 9.84% higher than the MF result respectively, which shows that the performance of NWS is better. Therefore, NWS was applied as the text-based method in TBN.

Additionally, according to the literature in [12–19], we provide a conclusion comparison in Table 2. The triangle symbols ('△' and '▲') in this table represent the ability and performance of a method, but the bold triangle ('▲') means its performance is the best. The symbol 'x' represents that there is none of this feature in the method. For example, Figure 4 shows a Swiss cheese model to express the processes of barrier failure to cause a derailment, in which the relations are linear. While, Figure 5 demonstrates the practical non-linear relationship of barriers. Briefly, Figure 5 shows several advantages as follows, compared with the Swiss cheese model in Figure 4: (1) shows that some barriers (e.g., B2 and B3) work together instead of in a strict sequence, (2) could express a parallel relationship such as B4a and B4b, and (3) can skip a barrier when its previous barrier fails. Thus, TBN is improved based on the graphical model to track the nonlinear relations. Therefore, compared with the Swiss model and TBN, they both have the character of linearity but TBN could have the character of nonlinearity; text data are adopted to provide objective prior data for the assessment model in TBN, so it shows fast and efficient compared to BN, because the priors in BN are from the expert's experience. In summary, TBN performs better than the Swiss model, fault tree and Bayesian network.

TBNs are improved based on the graphical model to track the nonlinear relations. Therefore, when comparing the Swiss cheese model and TBN, they both show a linear characteristic. However, the TBN also shows a nonlinear characteristic. As discussed in Table 2, text data are adopted to provide objective prior data for the assessment model in a TBN; therefore, it is faster and more efficient than the BN because the prior input in the BN is from the expert's experience. Overall, the TBN has fast and highly proficient properties and, compared to the BN, has more weight in terms of linear, nonlinear and dynamic characteristics; thus, the TBNs are represented by bold triangles.

Table 2. Summary of the comparison of text-based Bayesian network (TBN) with other methods.

Feature	TBN	Swiss Model	Fault Tree	Bayesian Network
Linear	▲	△	△	△
Nonlinear	▲	×	×	△
Dynamic properties	▲	×	×	△
Shows the relationships between barriers	▲	△	△	△
Text-based model	▲	×	×	×

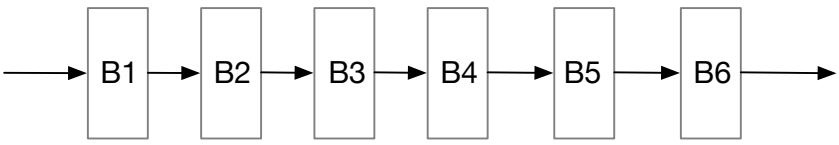


Figure 4. Swiss model for train derailment (please refer to Tables 1 and 3 for the meanings of the symbols).

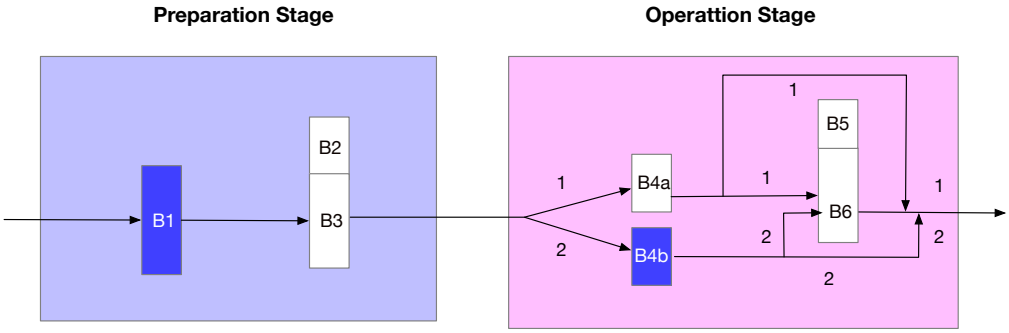


Figure 5. An illustration of the graphical model (please refer to Tables 1 and 3 for the meanings of the symbols).

Table 3. Train derailment for the different failures and their safety layers.

Symbols	Meanings
AS1	Preparation stage for 'inner failure'
AS2	Operation stage for 'inner failure'
BS1	Preparation stage for 'outer failure'
CS1	Preparation stage for 'both failures'
CS2	Operation stage for 'both failures'
T1	Successful derailment of 'inner failure'
T2	Successful derailment of 'outer failure'
T3	Successful derailment of 'both failures'

4.1. The Main Application of the TBN

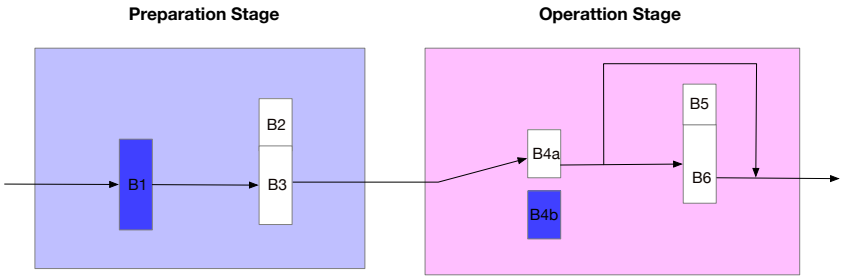
4.1.1. Graphical Model Establishment

In Section 3, Figure 5 shows a general graphical model of the derailment process. For the sake of simplicity, the graphical model in Figure 5 is divided into three basic graphical models under different scenarios. For example, in Figure 6, (a) considers only internal failures, (b) considers only external failures and (c) considers both internal and external failures. The relations between all barriers change with the failure type.

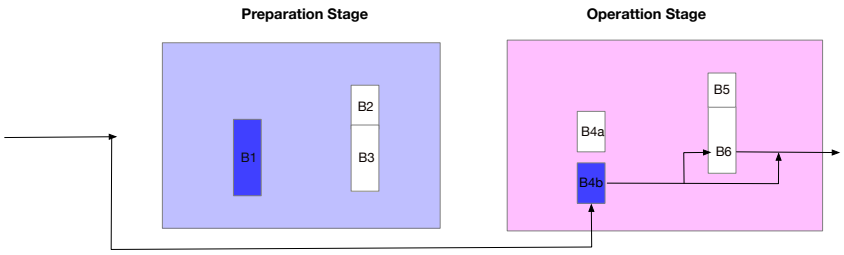
4.1.2. Fault Tree Establish

Since graphical models have been established, the relations between barriers have become clear. As shown in Figure 1, fault trees provide the prior probabilities. To further examine the graphical models depicted in Figure 7, we know that the prior probabilities of B1 and B4 can be determined from the proposed fault trees, but for certain barriers, such as B2, B3, B5 and B6, only partial probabilities are obtained from fault trees because they are related to other barriers, which means that conditional probabilities apply to those barriers.

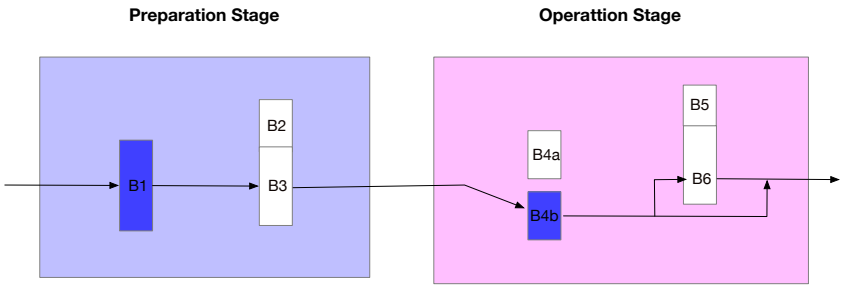
With the use of the proposed method, we first perform preprocessing, such as word cutting, feature extraction and feature analysis. By followings the steps mentioned in Section 3 to obtain fault trees, result-fault-trees are established, as shown in Figure 7. For example, B1 (the initial potential risk barrier) consists of two main causes (human and self-aging) in the first layer of the fault tree, which results from FT1. Moreover, they continuously change into several types, as shown in the third layer group, based on the text records. Then, the bottom layer contains the basic cause according to the text records. The third layer provides the prior probabilities for the BN, which is established in the following steps. In the fault trees shown in Figure 7, the basic logic relations are OR gates, and the root event of the fault trees could be better calculated. Figure 7 shows the structures of fault trees B1 and B3. It is observed that the second and third layers are similar except for the bottom layers.



(**a**) considers only internal failures



(**b**) considers only external failures

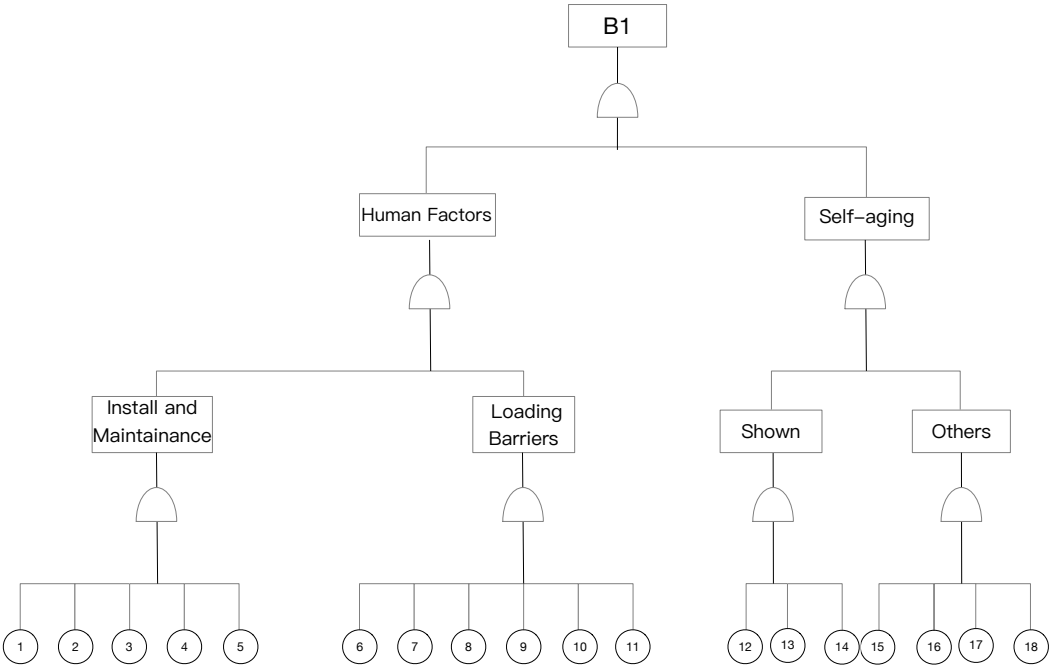


(**c**) considers both internal and external failures

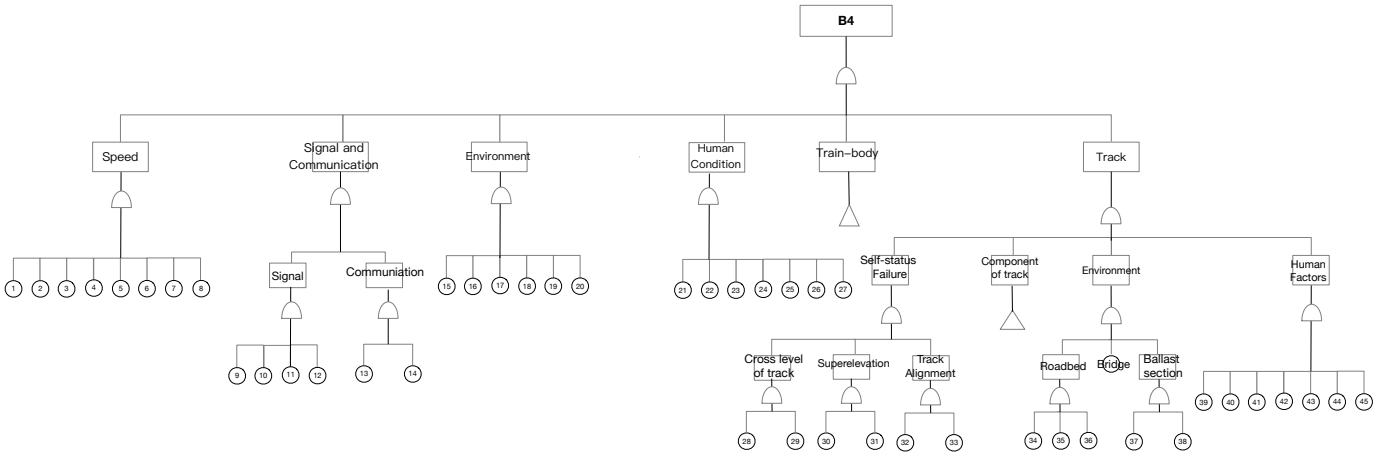
Figure 6. Graphical models under the different scenarios (please refer to Tables 1 and 3 for the meanings of the symbols).

4.1.3. Bayesian Network Establishment

Text records are extracted, and certain groups that are regarded as the basic fault trees are determined. To improve these basic fault trees, the process of accident determination based on text record features is proposed. Then, the basic fault trees are updated, and the railway accident graphical model is constructed (see Figure 7). Many barriers belonging to different fault trees are involved in the accident process. The graphical model (as shown in Figures 6 and 7) is then mapped into the BN. Additionally, during the process, the graphical and BN models are constructed. Based on the results of the fault trees and graphical model expressing the process of the train derailment accident, the BN is subsequently established (see Figure 8). In the BN, we define the nodes as causal factors and target events and the arcs as the dependence between the nodes. Figure 8 shows the BN model with two different layers and six barriers under three different scenarios. It also clearly demonstrates the barriers and layers.



(**a**) Fault tree of Barrier 1



(**b**) Fault tree of Barrier 4

Figure 7. Proposed fault trees from text data (please refer to the Appendix (Tables A1 and A2) for the meanings of the symbols).

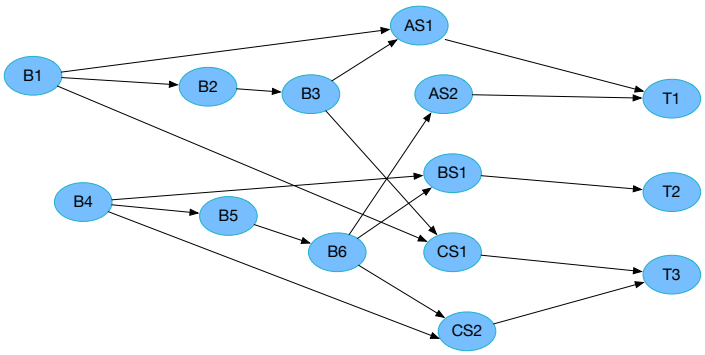


Figure 8. Proposed Bayesian network model (please refer to Tables 1 and 3 for the meanings of the symbols).

4.2. Quantitative Safety Assessment of the Railway System Using a Bayesian Network

The proposed BN model quantitatively represents the relevant dependencies. With the graphical model established and examined, CPTs are also compiled. Table 4 summarizes the prior probabilities of the barriers at risk. As such, the analysis results could reflect the practical transportation situation.

The established BN model based on the graphical model shown in Figure 6 describes train derailment accidents. The calculation results of the BN model will be used to assess the safety level of each operation scenario. In this section, the barrier safety potentials are analyzed considering the prior probabilities of the barriers at risk. Therefore, the barrier safety potentials are dynamically assessed under the various scenarios.

Table 4. Prior probabilities of the basic events.

Code	Definition	Count	Prior Probabilities
T110	Rack, Roadbed and Structures; Wide gage (due to defective or missing crossties)	1309	0.1645%
H702	Switch improperly lined	1355	0.1703%
T314	Switch point worn or broken	600	0.0754%
T207	Broken Rail—Detail fracture from shelling or head check	542	0.0681%
H307	Shoving movement, man on or at leading end of movement, failure to control	524	0.0646%
...	...	...	...

4.2.1. The Probabilities of Train Derailment

A train derailment refers to the occurrence of a derailment accident. The train derailment and failure probabilities of the safety layers are calculated using the established BN model depicted in Figure 8, and the results are provided in rows 2 and 5–7 of Table 5. Table 5 indicates the failure probabilities of the safety layers. Referring to the assessment criteria, we find that in Situation 3, high probabilities are attained that are mainly caused by the high failure probabilities in the first stage. In contrast, in Situation 2, the lowest success probabilities occur because both the first and second stages have low failure probabilities. It should be noted that the prior probabilities of the barriers originate from the text records, in contrast to the majority of previous works, which obtained prior probabilities from expert experience, thus resulting in uncertainty; this should be further examined.

It is worth mentioning that the probabilities in Table 5 are conditional probabilities, under the precondition that the derailment accident occurred through the train process. A comparison of the results under the different accident scenarios in Table 5 reveals the importance of considering the various potential derailment risks in the safety assessment process. If a safety manager focuses only on the prevention of the first stage, the safety level is considered acceptable (**Assessment criteria for the safety level** in Table 6) with a train derailment probability of 2.24×10^{-4}.

Table 5. The probabilities of actual accidents and safety layer failure.

	1	2	3
Prior probabilities of successful derailment	3.00×10^{-3}	1.31×10^{-1}	3.06×10^{-2}
Posterior probabilities of successful derailment (give successful safety barriers)	4.00×10^{-3}	1.30×10^{-2}	–
Failure probabilities of first stage	2.71×10^{-1}	1.30×10^{-2}	2.71×10^{-2}
Failure probabilities of second stage	1.30×10^{-2}	1.00	1.31×10^{-1}

Table 6. Assessment criteria for the safety level.

Safety Ability	Train Derailment Probabilities
Acceptable	$<1.00 \times 10^{-3}$
Tolerable	$[1.00 \times 10^{-3}, 1.00 \times 10^{-2}]$
Unacceptable	$>1.00 \times 10^{-1}$

4.2.2. Posterior Probability to Assess the Train Derailment Accident

By integrating different scenarios in the BN model, the derailment information obtained under one scenario could be applied to update the train derailment probabilities of the other scenarios. For example, when a failure actually occurs, the BN model is updated, and the posterior train derailment probabilities of the other three scenarios are provided in row 3 of Table 5. By comparing rows 2 and 3 of Table 5, it is observed that the updated train derailment probabilities have changed. This occurs because when train derailment occurs, barriers B1 and B2 are believed to have higher probabilities in the insecure state than previously estimated, and this change increases the failure probabilities and further enhances the train derailment probability under the other scenarios.

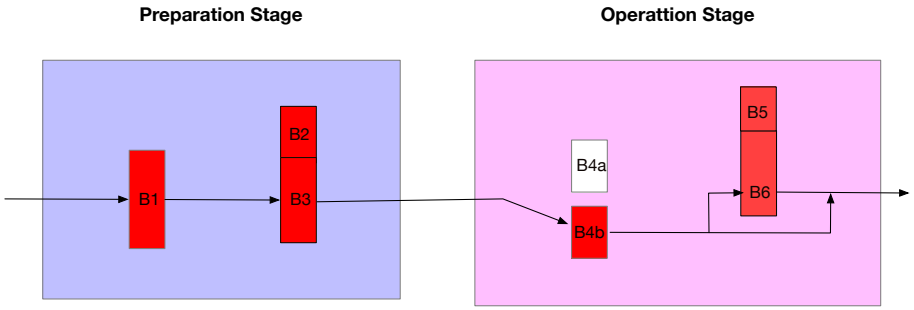
4.2.3. Dynamic Probability Assessment Given a Train Derailment Accident

Recently, BN models have been frequently applied to dynamically update the probability. Thus, the occurrence probabilities of barriers at risk may change over time. In this section, a BN model is adopted to diagnose the change in the barriers at risk based on the available evidence. Then, the accident probabilities are updated according to the posterior probabilities of the barriers at risk.

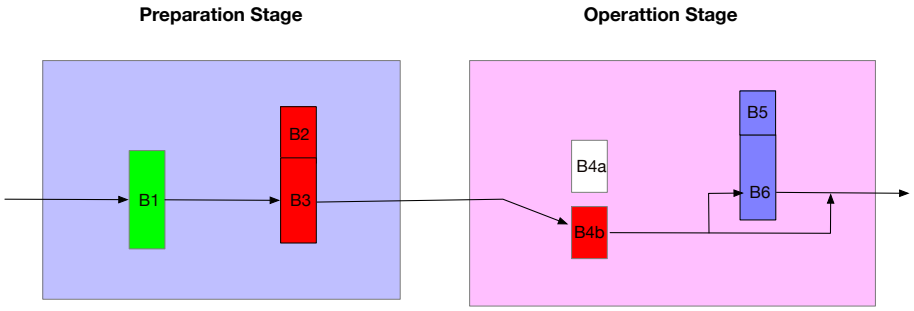
Figure 9 shows the posterior occurrence probabilities of the barriers at risk. According to Tables 6 and 7, it is observed that the safety potentials of B1, B5 and B6 have changed. Given the evidence, the BN model is updated. It is observed that Situation 1 exhibits growth when provided with successful safety barriers. With the use of solid evidence to conduct a dynamic assessment, the BN model provides a more reliable assessment of the defensive ability. Specifically, B1 is assigned green, while B4 and B5 are assigned purple. Thus, when evidence is used in the BN model, the posterior safety potentials of the barriers are obtained, which supports a more reliable weakness identification of the safety system. After the update, it is believed that B4 also exhibits a very low safety potential under one of the critical accident scenarios—accidents caused by employees. Therefore, in addition to B2, B3 and B4 are weak links in the safety system.

Table 7. Quantitative Evaluation.

Qualitative Evaluation	Quantitative Evaluation	Expression
Certain	1	Black
Very High	1.00×10^{-1}	Red
High	1.00×10^{-2}	Purple
Moderate	1.00×10^{-3}	Green



(**a**) Graphical model when considers only external failures



(**b**) Updated graphical model when considers only external failures

Figure 9. Updated graphical model with the safety potential under each scenario (please refer to Table 7).

4.2.4. Cause Chain Analysis with the Bayesian Network

As a result that the BN proposed in this paper is based on fault trees, i.e., the research is performed on limited text records, through the fault tree of Barrier 4, the fault tree causing B4 is identified, and by connecting the BN to the FT chain, a more consistent cause chain of the actual situation is determined. The most direct factors also can be quickly identified. For example, first, T3 is defined as a failure, and we can obtain $P(T3 = 1) = 0$. To acquire the cause chain, we should identify the parent nodes of T3 and choose the higher posterior probabilities among these parent nodes. Furthermore, the next parent nodes are assessed until posterior probabilities are no longer available. The following is a general example of how to determine cause chains.

1. $P(T3 = 1)$.
2. Identify the parent nodes of T3 to compare their posterior probabilities and choose the higher probability. CS1 is obtained, and we now have CS1→T3, while B1→CS1→T3 is acquired.
3. Repeat step (2) to identify the corresponding parent nodes with high probabilities, i.e., B4→CS2→T3.

Specifically, when a railway derailment accident occurs, the most likely cause involves the external factors during the driving process caused by B4. A more specific situation combined with the fault tree of B4 indicates that the B4 fault was caused by T110. In the fault tree of B4, when B4 is adopted as a safety barrier, the reasons for its failure include external factors, weather conditions and HFs. The external factors primarily include the rails and turnouts. "T110 Wide" represents gauge factors (due to defective or missing crisscrosses), so when the train is in operation, T110 is the main factor causing train derailment, and the specific cause chain is T110→ B4→CS2→T3.

5. Conclusions

This study mainly proposed a method to establish the TBN model where its prior probabilities are determined from text records of rail accidents. The established TBN model can quantitatively and accurately evaluate the safety situations of railways. The main

outcomes of the method we proposed are as follows: (1) Posterior probability of the train derailment accident. It can be revealed that if a safety manager focuses only on the prevention of the first stage, the safety level is considered acceptable with a train derailment probability of 2.24×10^{-4}. (2) Diagnose the change in the barriers at risk based on criteria for the safety level. Given by the assessment criteria, the result of dynamic probability assessment shows that B4 exhibits a very low safety potential under one of the critical accident scenarios—accidents caused by employees. B4 is the weak links in the safety system. (3) Give a key factor and obtain a cause chain within the key factor. For example, "T110 Wide" represents gauge factors (due to defective or missing crisscrosses), and when the train is in operation, T110 is the main factor causing train derailment, and the specific cause chain is T110→B4→CS2→T3.

The analysis results represent not only quantitative risk results but also the key barriers and critical potential risks of defects, including cause chains. Compared to previous works, whose prior probabilities are determined based on expert experience, this paper considers 10-year text record data. In this paper, the prior probabilities are acquired from text records, which helps improve the assessment accuracy and efficiency. Additionally, this paper demonstrates the accident occurrence principle and process in graphical form, which could help managers better understand and prevent accidents. The method establishes the dependency of the different parts. Hence, it updates the state of a given subsystem with evidence from other subsystems. This ensures that managers obtain the latest state information to the greatest extent, thus facilitating risk reduction. The work demonstrates relevant nonlinear relationships to improve the assessment results. Additionally, this paper studies the barrier safety potential to identify high-risk barriers, which can guide managers to better prioritize these barriers.

In future work, a more intelligent text analysis method will be employed in the first stage of the proposed method. Additionally, event trees will be adopted to address the weak links within the safety system and provide more references for railway management.

Author Contributions: Conceptualization, K.L. and F.K.; methodology, L.Y. and G.S.; software, L.Y.; formal analysis, L.Y.; writing—original draft preparation, L.Y.; writing—review and editing, L.Y., G.S., K.L. and F.K.; supervision, K.L. and F.K.; funding acquisition, K.L. and L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Fundamental Research Funds for the Central Universities (2018YJS204), the Beijing Natural Science Foundation (Grant No. 8202039) and the National Natural Science Foundation of China (Grant No. 71942006).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used during the study are available online (http://www.fra.dot.gov).

Acknowledgments: This work was supported by the Fundamental Research Funds for the Central Universities (2018YJS204), the Beijing Natural Science Foundation (Grant No. 8202039) and the National Natural Science Foundation of China (Grant No. 71942006).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The following symbols are used in Figure 7 (the text records from the Federal Railroad Administration (FRA) (http://www.fra.dot.gov)):

Table A1. The following symbols are used in Figure 7a.

Symbols	Meanings in the Text Records
1	Human Factor—Signal—Train Control—Installation or maintenance error (shop).
2	Computer system configuration/management error (non vendor)
3	Human Factor - Signal installation or maintenance error (field)
4	Improper train make-up at initial terminal
5	Computer system configuration/management error (vendor)
6	Overloaded car
7	Improperly loaded car
8	Trailer or container tiedown equipment improperly applied
9	Overloaded container/trailer on flat car
10	Improperly loaded container/trailer on flat car
11	Hazardous materials regulations, failure to comply
12	Broken rim
13	Damaged flange or tread (build up)
14	Journal (roller bearing) failure from overheating
15	Improper train inspection
16	Failure to comply with failed equipment detector warning or with applicable train inspection rules.
17	beyond checklist
18	Failure to detect minor by regular inspection

Table A2. The following symbols are used in Figure 7b.

Symbols	Meanings in the Text Records
1	Coupling speed excessive
2	Switching movement, excessive speed
3	Train on main track inside yard limits, excessive speed
4	Train outside yard limits, in block signal or interlocking territory, excessive speed
5	Failure to comply with restricted speed in connection with the restrictive indication of a block or interlocking signal.
6	Train outside yard limits in non block territory, excessive speed
7	Failure to comply with restricted speed or its equivalent not in connection with a block or interlocking signal.
8	Speed, other (Provide detailed description in narrative)
9	Power device interlocking failure
10	Power switch failure
11	Automatic cab signal displayed false proceed
12	Block signal displayed false proceed
13	Failure to stop train in clear
14	Radio communication equipment failure
15	Snow, ice, mud, gravel, coal, sand, etc. on track
16	Extreme environmental condition—TORNADO
17	Extreme environmental condition—FLOOD
18	Extreme environmental condition—DENSE FOG
19	Extreme environmental condition—EXTREME WIND VELOCITY
20	Other extreme environmental conditions (Provide detailed description in narrative)
21	Impairment of efficiency or judgment because of drugs or alcohol
22	Incapacitation due to injury or illness
23	Employee restricted in work or motion
24	Employee asleep
25	Employee physical condition, other (Provide detailed description in narrative)
26	Failure by non railroad employee, e.g., industry employee, to control speed of car using hand brake
27	other
28	Cross level of track irregular (at joints)
29	Cross level of track irregular (not at joints)
30	Superelevation improper, excessive, or insufficient
31	Superelevation runoff improper
32	Track alignment irregular (buckled/sunkink)
33	Track alignment irregular (other than buckled/sunkink)
34	Roadbed settled or soft
35	Washout/rain/slide/flood/snow/ice damage to track
36	Other roadbed defects (Provide detailed description in narrative)
37	Disturbed ballast section
38	Insufficient ballast section
39	Human Factor—track
40	Instruction to train/yard crew improper
41	Shoving movement, absence of man on or at leading end of movement
42	Shoving movement, man on or at leading end of movement, failure to control
43	Failure to couple
44	Manual intervention of classification yard automatic control system modes by operator
45	Humping or cutting off in motion equipment susceptible to damage, or to cause damage to other equipment

References

1. Heinrich, H.W. Industrial Accident Prevention. A Scientific Approach. In *Industrial Accident Prevention. A Scientific Approach*; McGraw-Hill Book Company: New York, NY, USA, 1941.
2. Underwood, P.; Waterson, P. Systems thinking, the Swiss Cheese Model and accident analysis: A comparative systemic analysis of the Grayrigg train derailment using the ATSB, AcciMap and STAMP models. *Accid. Anal. Prev.* **2014**, *68*, 75–94. [CrossRef] [PubMed]
3. Dindar, S.; Kaewunruen, S.; An, M.; Sussman, J.M. Bayesian Network-based probability analysis of train derailments caused by various extreme weather patterns on railway turnouts. *Saf. Sci.* **2018**, *110*, 20–30. [CrossRef]
4. Li, K.; Wang, S. A network accident causation model for monitoring railway safety. *Saf. Sci.* **2018**, *109*, 398–402. [CrossRef]
5. Xin, M.; Ke-Ping, L.; Zi-Yan, L.; Jin, Z. Analyzing the causation of a railway accident based on a complex network. *Chin. Phys. B* **2013**, *23*, 028904.
6. Zhou, J.; Xu, W.; Guo, X.; Ding, J. A method for modeling and analysis of directed weighted accident causation network (DWACN). *Phys. A* **2015**, *437*, 263–277. [CrossRef]
7. Zhou, J.; Xu, W.; Guo, X.; Liu, X. A hierarchical network modeling method for railway tunnels safety assessment. *Phys. A* **2017**, *467*, 226–239. [CrossRef]
8. Liu, J.; Schmid, F.; Zheng, W.; Zhu, J. Understanding railway operational accidents using network theory. *Reliab. Eng. Syst. Saf.* **2019**, *189*, 218–231. [CrossRef]
9. Liu, P.; Yang, L.; Gao, Z.; Li, S.; Gao, Y. Fault tree analysis combined with quantitative analysis for high-speed railway accidents. *Saf. Sci.* **2015**, *79*, 344–357. [CrossRef]
10. Dindar, S.; Kaewunruen, S.; An, M.; Gigante-Barrera, Á. Derailment-based fault tree analysis on risk management of railway turnout systems. In *IOP Conference Series: Materials Science and Engineering*; IOP Science: Prague, Czech Republic, 2017; Volume 245, p. 042020.
11. Lorenc, A.; Kuźnar, M.; Lerher, T.; Szkoda, M. Predicting the Probability of Cargo Theft for Individual Cases in Railway Transport. *Teh. Vjesn.* **2020**, *27*, 773–780.
12. Song, G.; Khan, F.; Wang, H.; Leighton, S.; Yuan, Z.; Liu, H. Dynamic occupational risk model for offshore operations in harsh environments. *Reliab. Eng. Syst. Saf.* **2016**, *150*, 58–64. [CrossRef]
13. Van Staalduinen, M.; Khan, F. A barrier based methodology to assess site security risk. In *SPE E&P Health, Safety, Security and Environmental Conference-Americas*; Society of Petroleum Engineers: Denver, CO, USA, 2015.
14. Chen, R.; Zhu, S.; Hao, F.; Zhu, B.; Zhao, Z.; Xu, Y. Railway Vehicle Door Fault Diagnosis Method with Bayesian Network. In Proceedings of the 2019 4th International Conference on Control and Robotics Engineering (ICCRE), Nanjing, China, 20–23 April 2019; pp. 70–74.
15. Rathnayaka, S.; Khan, F.; Amyotte, P. SHIPP methodology: Predictive accident modeling approach. Part I: Methodology and model description. *Process Saf. Environ. Prot.* **2011**, *89*, 151–164. [CrossRef]
16. Rathnayaka, S.; Khan, F.; Amyotte, P. SHIPP methodology: Predictive accident modeling approach. Part II. Validation with case study. *Process Saf. Environ. Prot.* **2011**, *89*, 75–88. [CrossRef]
17. Młynarski, S.; Pilch, R.; Smolnik, M.; Szybka, J. Methodology of network systems reliability assessment on the example of urban transport. *Eksploat. Niezawodn.* **2018**, *20*. [CrossRef]
18. Adedigba, S.A.; Khan, F.; Yang, M. Dynamic safety analysis of process systems using nonlinear and non-sequential accident model. *Chem. Eng. Res. Des.* **2016**, *111*, 169–183. [CrossRef]
19. Song, G.; Khan, F.; Yang, M. Security assessment of process facilities–Intrusion modeling. *Process Saf. Environ. Prot.* **2018**, *117*, 639–650. [CrossRef]
20. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
21. Carretero-Campos, C.; Bernaola-Galván, P.; Coronado, A.; Carpena, P. Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A* **2013**, *392*, 1481–1492. [CrossRef]
22. Kuhn, T.; Perc, M.; Helbing, D. Inheritance patterns in citation networks reveal scientific memes. *Phys. Rev. X* **2014**, *4*, 041036.
23. Beliga, S.; Martinčić-Ipšić, S. Node selectivity as a measure for graph-based keyword extraction in Croatian news. In Proceedings of the 6th International Conference on Information Technologies and Information Society (ITIS2014), Šmarješke Toplice, Slovenija, 2014.
24. Yang, L.; Li, K.; Huang, H. A new network model for extracting text keywords. *Scientometrics* **2018**, *116*, 339–361. [CrossRef]
25. Yang, L.; Li, K.; Zhao, D.; Gu, S.; Yan, D. A Network Method for Identifying the Root Cause of High-Speed Rail Faults Based on Text Data. *Energies* **2019**, *12*, 1908. [CrossRef]