Construction and Building Materials 346 (2022) 128353

Contents lists available at ScienceDirect

Construction and Building Materials

journal homepage: www.elsevier.com/locate/conbuildmat

A machine learning based methodology for broken rail prediction on freight railroads: A case study in the United States

Xin Wang [a], Xiang Liu [a,*], Zheyong Bian [b,1]

[a] Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[b] Department of Construction Management, University of Houston, Texas, USA

ARTICLE INFO

Keywords:
Broken rail
Predictive analytics
Machine learning
Freight railroads

ABSTRACT

Predicting the occurrence of broken rails has safety and economic benefits, and reduces accidents and service disruptions. This paper aims to build a data-driven model for broken rail prediction using data related to infrastructure, operations, inspections, and weather conditions from 2013 to 2019. The railroad data was provided by one major Class I U.S. railroad. The weather condition data was collected from the National Oceanic and Atmospheric Administration (NOAA). Based on time-series data partitioning, three different machine learning models are developed for predicting broken rail occurrence one month in advance. The selected models, including logistic regression, random forests, and gradient boosting, are trained using data from 2013 to 2018. The performance of three trained models are evaluated using the data from 2019. The relationship between the percentage of the network scanned and the percentage of broken rails found is used to identify locations that are more prone to broken rails. The findings of this study show that the gradient boosting model performs better than the other two methods for our datasets. The model also identifies that the number of detected rail defects within last 365 days, minimum ambient temperature of the last 30 days, days from the last broken rail, segment length, traffic density and other factors have significant influences on prediction results. Using this model, 40.4% of broken rails can be successfully predicted one month in advance when focusing on 10% of the railroad network scanned. The model can potentially be used to prioritize inspection and maintenance activities on broken-rail-prone locations, and thus to further improve infrastructure safety given budgetary constraints.

1. Introduction

Broken rails (aka rail service failures) are the leading cause of severe freight-train accidents in the United States [1,2]. For example, there were 219 broken-rail-caused freight train derailments on one mainline of a Class I railroad[2] from 2000 to 2014, accounting for 22% of all types of derailments, exceeding all other causes [3]. Annually, Class I freight railroads incur about $83 million in damage costs and service interruptions due to broken rail occurrences [4].

Improvements in rail manufacturing and inspection technology have significantly reduced the incidence of broken rails [5,6]. In our studied data provided by one Class I railroad in the United States, the average annual number of broken rails has decreased from 1,940 two decades ago (2000–2009) to 806 in the last decade (2010–2019). However, a number of broken rails still occur. Efforts are underway in the industry to keep this momentum and further reduce broken rail risk. While additional inspection and/or maintenance would be helpful for risk reduction, it would require additional financial resources. Therefore, knowing how to allocate limited resources to maximize safety benefits is of great interest and practical value.

This paper focuses on broken rail occurrence prediction using a machine learning based methodology. The proposed methodology is applied in practice based on the multi-source data (infrastructure, operation, inspection) provided by a major freight railroad company in the United States, as well as climatological data from the NOAA. Based on time-series data partitioning, three different machine learning

* Corresponding author at: Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, 500 Bartholomew Road, Room 428D, Piscataway, NJ 08854, USA.

E-mail address: xiang.liu@rutgers.edu (X. Liu).

[1] This work was done when Dr. Zheyong Bian was at Rutgers University.

[2] Class I railroad was defined in 1992 by the Surface Transportation Board [7] as being any carrier earning annual revenue greater than $250 million. This has since been adjusted for inflation and was most recently set to $504,803,294 in 2019.

https://doi.org/10.1016/j.conbuildmat.2022.128353
Received 23 January 2022; Received in revised form 27 April 2022; Accepted 2 July 2022
Available online 13 July 2022
0950-0618/

Table 1
Comparison of Selected Data-Driven Models of Broken Rail Prediction.

Objective	Scope of data	Methodology	Model variables	Data partitioning and model performance	Authors and research
Estimate occurrence of broken rails during a two-year period	3,676 records from one U.S. railroad, 1998–2000	Logistic regression model	Rail characteristics, traffic-related variables, curve degree, speed, presence of turnout, and combinations of variables	Accuracy: 87.4%	Dick (2001) [8]; Dick et al. (2002) [9]; Dick et al. (2003) [10]
Estimate occurrence of broken rails during a four-year period	15,999 records from one U.S. freight railroad, 2003–2006	Hybrid logistic regression and neural network hybrid model	Rail characteristics, traffic-related variables, speed, curvature, superelevation, grade, tie work, grinding, presence of turnout, presence of rail defect and geometry defect, grinding, presence of infrastructure, and combinations of variables	Testing accuracy: 67.9%	Schafer and Barkan (2008) [11]
	25,370 records from one U.S. freight railroad, 2003–2006	Logit model	Rail characteristics, traffic-related variables, presence of rail defect and geometry defect, and presence of bridge	Random partitioning: 60% of data for training, 40% of data for testing Testing accuracy: 66.3%	Schafer and Barkan (2008) [4]
Predict the frequency of broken rails	–	Fuzzy logit model	Tonnage, temperature, rail age, and curve degree	Predicted number of broken rails in 2010 is 6.72 and real number is 6.	Vesković et al. (2012) [12]
Estimate rail service life	–	Weibull distribution	Rail characteristics, curve, season (winter or summer), tonnage	–	Chattopadhyay and Kumar (2009) [13]
	648 records from freight and passenger railroad, 2010–2015	Markov stochastic processes	Tonnage, rail defects, curve, grade, segment length	The first quartile and third quartile errors are −1.8 and 0.8 years	Bai et al. (2017) [16]
Predict the risk of broken rails between two rail inspections	290,735 records (21,000 miles) freight railroad, 2011–2016	Survival analysis model	Tonnage, inspection schedules, segment length, broken rails, rail defects, geometry defects, days from last grinding, grade, curve degree, rail age, product age curvature, and turnouts	Training data from 2011 to 2015, testing data in 2016 Testing accuracy: 68.3%	Ghofrani et al. (2020) [17]
Predict the occurrence of broken rails between two rail inspections	5,270 miles from heavy haul line 2011–2016	Gradient boosting	Tonnage, rail age, temperature, segment length, rail inspection schedules, rail defects, geometry defects, days from last grinding, grade, curve degree, turnouts, product of age curvature	G-Mean: 0.95, AUC: 0.95	Ghofrani et al. (2021) [2]

models, including logistic regression, random forests, and gradient boosting, are developed and trained using the data from 2013 to 2018. The performance of three trained models is evaluated using the data from 2019. Additionally, factors that significantly influence broken rail prediction results are identified using the trained model. Finally, the case study shows that the model can potentially be used to prioritize inspection and maintenance activities on broken-rail-prone locations, and thus further improve infrastructure safety given budgetary constraints.

The organization of the article is as follows. Section 2 discusses the topic background and the significant contributions of the current study. In section 3, the methodology of the model in this study is demonstrated. Section 4 shows an overview of the data collected and the analysis of input variables of the models. Subsequently, the results of different models are compared. Discussions of the results and future directions of the study are provided in Section 5 and Section 6.

2. Literature review

Broken rails are mainly caused by fatigue, which can be initiated at the rail head, at the web, and at the foot [46]. There is an increased risk of broken rails in winter because tensile thermal stresses reach peak values in cold weather [47]. Previous research on data-driven analysis on broken rails falls into three categories: 1) Statistical models proposed for predicting the occurrence and frequency of broken rails [4,8–12]; 2) Machine learning based models for predicting the occurrence and frequency of broken rails [2,4,11]; 3) Probabilistic models used for analyzing the risk of broken rails under different scenarios [13–17].

Statistical models were developed to predict the occurrence and frequency of broken rails over a specific period. A multivariate statistical model based on logistic procedure was developed to predict the probability of broken rails over a two-year period [4,8–11]. Stepwise regression was used to determine the most relevant parameters and combinations of parameters for inclusion in the model. A probability threshold is determined to estimate whether a broken rail is predicted to occur. A fuzzy model based on the fuzzy logic process was developed to predict the frequency of broken rails [12]. This model was shown to perform well with basic factors that are relatively easy to collect for long segments. Nevertheless, it would not have an accurate prediction over a short segment due to data limitations.

The second type of data-driven analysis refers to machine learning models for predicting the occurrence and frequency of broken rails. Two types of hybrid model that combine a multilayer perception (MLP) neural network and logistic regression were proposed to predict the occurrence of broken rails over a four-year period, which outperformed the logistic regression model in terms of accuracy [4,11]. The first hybrid model applied a logistic regression model to pre-select the most useful variables, and then developed an MLP model based on these selected variables. The second hybrid model was developed using the logistic regression model to calculate the probability of broken rail and added the calculated probability as an additional input variable into the MLP model. More recently, a gradient boosting classifier with resampling and bootstrap aggregation (or bagging) incorporated into the algorithm was implemented to predict the occurrence of broken rails between two successive rail inspections using data from a railroad for the six-year period from 2011 to 2016 [2].

Probabilistic models were proposed to analyze the risk of broken rail impacted by various factors. Chattopadhyay and Kumar [13] applied

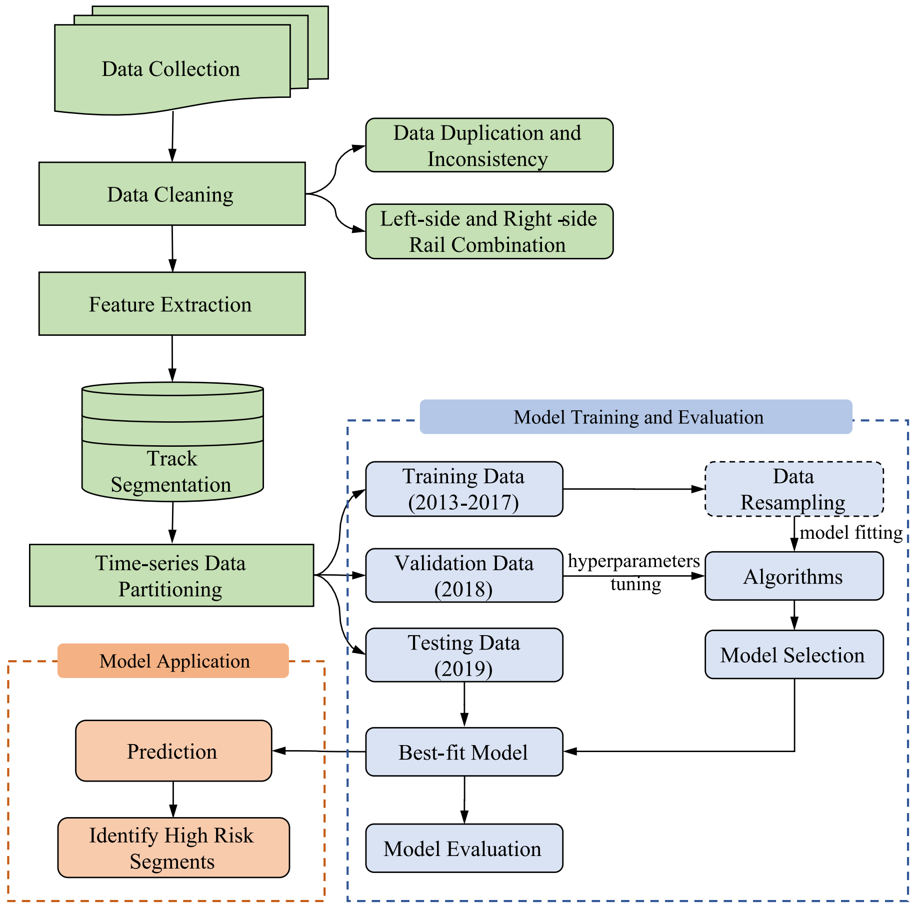

Fig. 1. Machine Learning Framework for Broken Rail Prediction.

Weibull distribution to model the relationship between the probability of broken rail and million gross tons (MGT), where the failure probability function monotonically increased with MGT. A defect-based risk analysis methodology for estimating broken rail risk was developed [14]. The Bayesian inference method was initially employed to capture a robust model of squat (over 4 mm) broken rail probability. Then, squat was divided into four experimental categories according to crack depth and visual length. Finally, relying on the broken rail probability scenarios and severity categories, the broken rail risk was defined. Along with directly using the crack information as the index for evaluating the squat severity, MGT was adopted as an intermediate variable for evaluating the growth length [15]. This is because rail staff can more easily collect the MGT data and apply the model in management. A model for estimating the rail service life of a section (1 km, 0.63 miles) was built using Markov processes, where the service life was defined as the period from the date when the rail is put into use to the date when the rail is replaced [16]. The estimated rail service life is close to the real rail service life. However, it assumed that if a severely defective rail was replaced and the length of the rail was small (for example, 25 m), such unplanned replacement activities would have little effect on the overall service life of the section, which means the proposed model cannot be applied to identify rail condition over a short segment. A survival analysis model was developed to predict the risk of broken rails between two successive rail inspections using multiple data sources [17]. This study also analyzed the impact of covariates on rail life defined by the total cumulative tonnage. Further details about the comparison of data-driven models for broken rail prediction are shown in Table 1.

Nevertheless, the problem of data-driven analysis of broken rails is still understudied in the current literature. Many prior related studies have focused on a relatively longer prediction period (e.g., one year) for railroad capital planning. Different from long-term prediction models, the development of a short-term prediction model for broken rails is more challenging due to the scarcity of broken rail occurrences in shorter periods. However, a short-term prediction model would be helpful in guiding inspection and maintenance [18,19]. Thus, in this paper, we develop an effective framework that is customized for broken rail prediction in a shorter period, specifically one month in advance. This framework consists of a pipeline of methodologies, including data cleaning, data partitioning, feature extraction, track segmentation, covariance analysis, and hyperparameter optimization. Additionally, compared to the literature using older datasets, we have collected and analyzed more recent data from one major freight railroad. Finally, in this research, as compared to the data partitioning performed in most previous studies, testing data for evaluating the performance of the proposed models were generated by partitioning all data according to time sequences to avoid data leakage. This paper provides the following contributions:

1. A framework consisting of a pipeline of methodologies is proposed for short-term prediction of broken rails. The effectiveness of the proposed framework is validated using data from one Class I railroad.
2. Feature-based track segmentation is proposed to delineate the railroad network, which is able to improve the performance of the proposed model by reducing variations in data records.
3. This paper extracts more features and uses them as input variables for the machine learning models. While previous studies did not include them, the feature importance identifies the top influencing variables, such as minimum temperature and crossing angles of turnouts, for broken rail prediction.

3. Methodology

A machine learning based methodology is proposed in this paper to predict the occurrence of broken rails. Fig. 1 shows a methodological framework for data-driven broken rail prediction. First, data cleaning is used to handle the issues of the raw data collected from one Class I U.S. railroad and the NOAA. Subsequently, important features (input variables) are extracted according to the processed data. Next, heterogeneous datasets are combined into an integrated dataset, and track segmentation is applied to combine shorter segments with the same or similar attributes into longer segments. By partitioning data according to time series, the data from 2013 to 2017 are defined as training data, and the data from 2018 and 2019 are separately selected as validation data and testing data. The model fitted by the training data (with and without data resampling) and the validation dataset is evaluated using testing data. Finally, the model is applied to a network-level case study on a major freight railroad.

3.1. Data cleaning

Due to human factors and possible errors from inspection machines, there can sometimes be more than one record associated with identical locations and time windows (if any), with observations being uniform (data duplication) or different (data inconsistency). Data cleaning aims to address data duplication and inconsistency issues and to process information combination for the left-side and right-side rail. To address the data duplication problem, only one of the duplicated records is kept in the database. To address the data inconsistency issue, the record with the worst condition information (to be conservative) will be retained in the database. For instance, if there is more than one record concerning curve information in the same segment, the maximum curve degree among different records would be assigned to this segment. Further, some databases differentiate between the left-side and right-side rail of the same track. In this study, data from two sides of the same track are combined according to their data types since many important datasets (e.g., traffic density) are recorded on the track level.

3.2. Data partitioning

When some data attributes are time-related and dynamically change with time, time-series splitting can provide a statistically robust model evaluation. If the testing data is randomly selected from all the original data, information regarding the future condition may leak into the model, which would lead to an over-optimistic estimation of the model. Therefore, in this study, time-series partitioning is used to sort all data into training, validation, and testing data. Specifically, the data from 2019 is selected as the testing data for evaluating model generalization. The data from 2018 is used as validation data which provides an unbiased evaluation of a model fit on the training data from 2013 to 2017 when model hyperparameters are tuned.

3.3. Data resampling

In the railroad network, the vast majority of track segments did not experience broken rails in a short time period, leading to imbalanced classification issues. As a result of this issue, the performance of a machine learning model leans partially toward the majority class in the imbalanced data. Data resampling aims to balance data by reducing the majority class samples or resampling the minority class samples, known as undersampling and oversampling, respectively [20–22]. Thus, data resampling is applied to address imbalanced data issues in this research.

Cluster-based majority undersampling is applied by generating a smaller set, based on centroids, using clustering methods. This algorithm samples the majority class by replacing a cluster of majority samples with the cluster centroid of a K-mean method, where the number of clusters is set according to the required level of undersampling [23]. Traditional oversampling algorithms, such as the synthetic minority oversampling technique (SMOTE), create synthetic examples in feature space from randomly selected pairs of feature examples from the minority class [24,25]. Based on SMOTE, He et al. [26] developed an adaptive synthetic sampling approach (ADASYN) by shifting the classifier decision boundary to focus more on those samples that are difficult to learn. It uses a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated. Therefore, in this research, cluster-based majority undersampling is used as an undersampling method, and ADASYN is the oversampling technique used in the machine learning models.

3.4. Prediction models

3.4.1. Logistic regression

Logistic regression is used to model the probability of the target object as a function of other variables. Consider a set of prediction vectors (feature vectors) $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ where n is the number of records, $\boldsymbol{x}_i = [k_1, k_2 \cdots k_d]$ is the i^{th} record of a column vector containing the values of d variables. The response variable for the column vector $\boldsymbol{x}_i$ is y_i where $y_i \in [0, 1]$. Logistic regression establishes the relationship between prediction vectors and the response variable as follows: $P(y|\boldsymbol{x}_i) = \frac{1}{1+e^{-(\boldsymbol{w}^T \boldsymbol{x}_i + b)}}$. Maximum likelihood estimation (MLE) is applied to estimate the parameters $\boldsymbol{w}$ and b that maximize the conditional likelihood of $\prod_{i=1}^{n} P(y_i | \boldsymbol{x}_i; \boldsymbol{w}, b)$.

3.4.2. Random Forests

Random Forests is one of the most famous bagging (Bootstrap Aggregating) algorithms: a bagged decision tree [27,28]. It does not require a lot of preprocessing and can be highly advantageous for training with heterogeneous data. Random Forests works according to the following steps:

Step 1: Sample m datasets $X_1, \cdots, X_m$ from all training data X.

Step 2: In each $X_j (j = 1, 2, \cdots, m)$, only a limited number (smaller than d, number of variables) of prediction vectors are used to train a full decision tree $h_j(X_j)$.

Step 3: The final performance of the Random Forests model $H(X)$ is evaluated by averaging all the decision tree models $H(X) = \frac{1}{m}\sum_{j=1}^{m} h_j(X_j)$.

3.4.3. Gradient boosting

Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models in an iterative fashion. In iteration t, the classifier $\alpha_t h_t(x)$ is added to the ensemble, where α_t is the weight of the classifier $h_t(x)$. Thus, all T classifiers are used to create the ensemble classifier $H_T(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$.

XGBoost is a scalable tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Traditional gradient boosting uses a gradient algorithm to minimize errors in the sequential model, while XGBoost is optimized by parallel processing, tree-pruning, and handling regularization to avoid overfitting [29,30]. Thus, XGBoost is ultimately the selected gradient boosting model in this study.

3.5. Evaluation metric

The evaluation metric is used to measure the performance of a machine learning model. Broken rail data only takes up a small proportion of the whole railroad network. Misclassifying observations of these territories has much more severe consequences, including service interruption or even derailment. A reasonable evaluation metric is able to provide a trade-off between the majority class and minority class. Typically, binary classification results can be divided into four categories (aka confusion matrix): the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The

Table 2
Distribution of Annual Tonnage of the Whole Network.

Average Annual Tonnage (MGT)	0–10	10–20	20–30	30–40	Over 40
Track Miles	8241	3344	3340	2030	3547
Percent of the Whole Network	40.20%	16.31%	16.29%	9.90%	17.30%

following measures can be obtained [31]:

True positive rate (recall or sensitivity):.$TPR = \frac{TP}{TP+FN}$

True negative rate (specificity):.$TNR = \frac{TN}{TN+FP}$

False positive rate:.$FPR = \frac{FP}{TN+FP}$

False negative rate:.$FNR = \frac{FN}{TP+FN}$

Positive predictive value (precision):.$PPV = \frac{TP}{TP+FP}$

To resolve the imbalanced data issue, oversampling the minority and using the geometric mean (G-Mean) as the evaluation metric were implemented for dealing with the same problem [2]. G-Mean computes the geometric mean of the accuracy of two classes. It maximizes the square root of the product of TPR and PPV, while obtaining a good balance. On the other hand, the receiver operating characteristics (ROC) curve and area under the ROC curve (AUC, $AUC = \frac{1+TPR-FPR}{2}$) provide a visualization and index of the relative trade-off between TPR and FPR.

The goal of the proposed model is to identify more broken rails with fewer track miles inspected. For a given score threshold T that determines the confusion matrix, $TPR(T)$ represents the percentage of broken rails detected, and the percentage of railroad network inspected is $P(T) = \frac{TP+FP}{TN+FP+FN+TP}$ if all segments have identical lengths. When it comes to an extremely imbalanced classification problem, as in this study, only thousands out of millions of total segments have broken rails. Thus, it is clear to have $TP \ll FP$ and $TP + FN \ll TN + FP$, and the $P(T)$ can be approximately equal to FPR. Finally, the customized evaluation metric E in this study is given by:

$$E = \int_0^1 TPRdP = \int_{-\infty}^{\infty} TPR(T)P^{'}(T)dT \approx \int_{-\infty}^{\infty} TPR(T)FPR^{'}(T)dT \quad (1)$$

3.6. Hyperparameter optimization

Underfitting and overfitting are two challenges associated with model training [32,33]. They are concerned with the degree to which the data in the training set is extrapolated to apply to unknown data. Overfitting refers to a phenomenon where the model learns a function with very high variance to perfectly model the training data, while underfitting occurs when the machine learning model is incapable of capturing the variability of the data [34]. The hyperparameter refers to the configuration of the model whose setting cannot be estimated from the training process. Appropriate hyperparameters provide a trade-off between bias and variance, and avoid the mistakes of underfitting and overfitting.

Bayesian optimization is selected as the hyperparameter optimization approach. On one hand, grid search and random search would require massive trials, especially for high dimensional data, which would affect the ultimate accuracy of model results. On the other hand, Bayesian optimization internally maintains a Gaussian process model of the objective function and uses objective function evaluations to train the model. One innovation in Bayesian optimization is the use of an acquisition function to determine the next point for evaluation [35]. The acquisition function can balance sampling at points that have low modeled objective functions, and explore areas that have not yet been modeled well.

4. Case study

The data collected for the current research comes from one Class I U. S. railroad for seven years from 2013 to 2019 and weather condition data from the NOAA. The railroad datasets provide infrastructure design, operation, and inspection information. The NOAA data includes climatological information for over 20,000 track miles.

4.1. Database description

4.1.1. Infrastructure data

Infrastructure data includes curvature data, grade data, rail data, turnout data, and signal data of the track. Curvature data indicates degree of the curvature, length of the curvature, length of spiral, superelevation, and direction of the curvature that shows the horizontal alignment of the track. Grade data (slope) shows the vertical alignment of the track. Rail data records the latest rail characteristics, including rail weight, the laid date of the rail, new versus re-laid rail, and joint versus continuous welded rails (CWR).

Turnout data specifies the turnout direction, frog type, turnout size, and other related information. In this study, the crossing angle is used so that the greater angle may incur more severe operation conditions [36]. The signal data indicates whether a track is in a signalized territory.

4.1.2. Operation data

Operation datasets contain the maximum allowable speed and traffic data. Speed data provides the maximum allowable speed of track on different segments. Traffic data specifies the track segments, corresponding gross tonnage, and gross number of passing cars. This study uses the monthly gross tonnage. Table 2 illustrates the distribution of the annual gross tonnage of the whole network.

4.1.3. Inspection related data

Broken Rail Data

Broken rails can be detected by the signal system, track inspectors, or maintenance of way crews in the field [8]. Once a broken rail is detected and the failure type is investigated, information regarding the broken rail is recorded in this database, such as the type of failure and found date. Ordinary broken rail (BRO) refers to a partial or complete break in which there is no sign of a fissure [49]. In our studied data, BRO is the primary type of broken rail, accounting for 28.15% of all broken rails. Furthermore, there is a significant difference among the average numbers of broken rails in different months (Fig. 2). For example, the average number of broken rails in January (197) is ten times greater than that in July (18). Broken rails in November, December, January, and February accounted for 64.3% of the total number of broken rails in our dataset.

Ultrasonic Rail Inspection Schedule Data

Ultrasonic rail inspection schedule data records the location, direction, and time when an inspection vehicle passes. The monthly average number of track miles inspected in this period is 9,010. The inspection interval of ultrasonic vehicles (in number of days) is shown in Fig. 3. The minimum inspection interval is 16 days. It also indicates that most segments are distributed in several inspection intervals such as 30 days, 60 days, 120 days, and 180 days. Fig. 4 demonstrates the cumulative percentage of the railroad network (non-repetitive) with ultrasonic rail flaw inspection. It shows that about 88.7% of the whole network track (non-repetitive) is inspected by an ultrasonic vehicle at least once every 200 days.

Rail Defect Data

Cyclic dynamic load on the rail can result in rail deterioration, and the rail can also be subject to mechanical and thermal forces during installation and track maintenance operations. Ultrasonic rail inspection is applied to detect rail defects, including the size and type of defects. About 78,000 rail defects (excluding broken rail) were found between 2013 and 2019 in the studied railroad network, with approximately 0.53 rail defects per mile per year. Detail fracture (TDD) is the dominant type of rail defect, accounting for 26.8% of all rail defects. TDD means a progressive fracture originating at or near the surface of the rail head

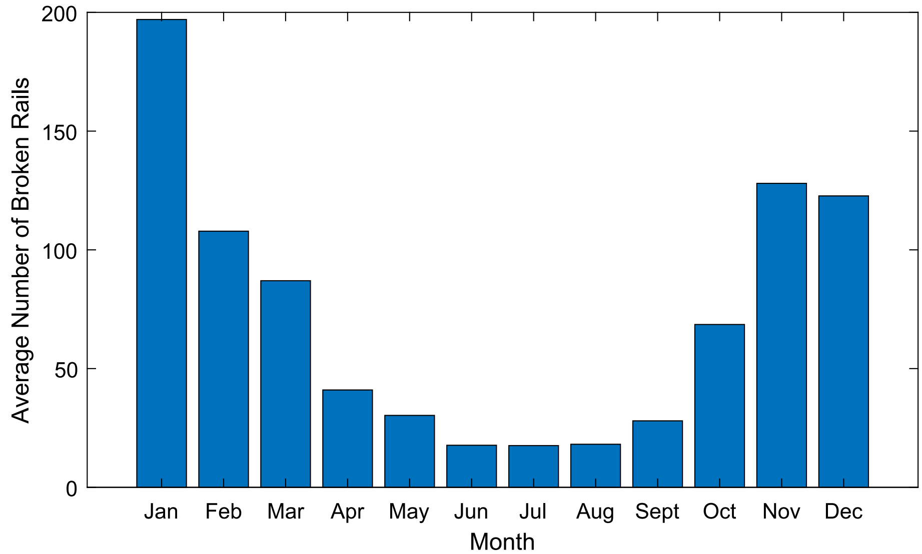

Fig. 2. Distribution of Average Monthly Number of Broken Rails on the Studied Railroad.

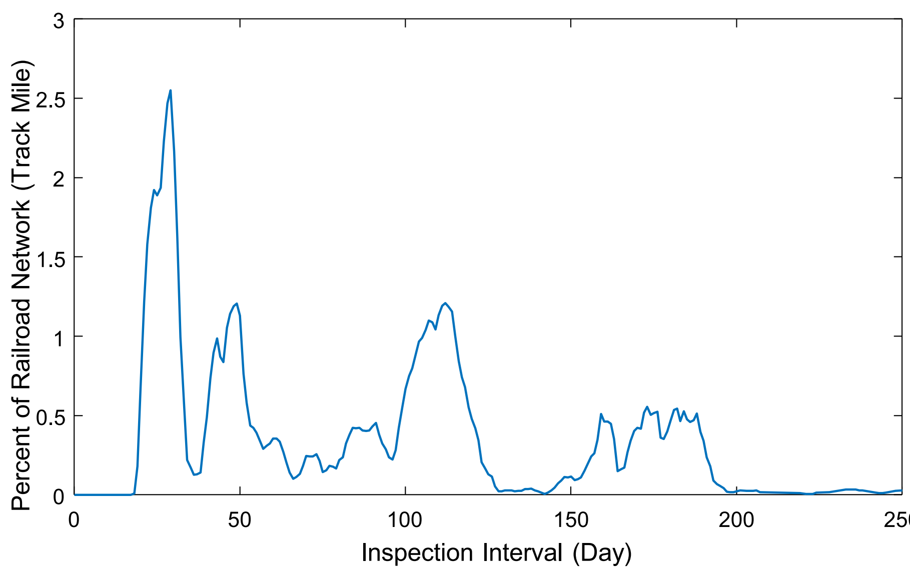

Fig. 3. Distribution of Ultrasonic Rail Flaw Inspection Interval.

[49]. It may originate from shelly spots, head checks, or flaking.

Track Geometry Exception Data

Track geometry refers to the geometrical data of the track, such as alignment, profile, cross-level, warp, and gauge. As the geometry deteriorates, the dynamic loads on the rail increase, resulting in the occurrence of rail defects or broken rails [37,38]. Once geometry defects (including wide gauge) were found and corrected if needed, types of geometry exception, size of the defects, found date, and other related data were recorded in the track geometry exception data. Among all the exception types, wide gauge is the primary type of track geometry defect, accounting for 16.1% of all defects. Wide gauge is one of the most important causes of derailments on main lines [50]. In this research, all types of geometry exception are included in the model.

Vehicle-Track Interaction Exception Data

The Vehicle Track Interaction (VTI) system is used to measure car body accelerations, truck frame accelerations, and axle accelerations [39–41]. It can assist in the early identification of vehicle dynamics that might lead to rapid degradation of track and equipment [48]. The VTI database includes the location information, exception date, exception type, exception value, and exception priority [51].

4.1.4. *Climatological data*

The NOAA collects local climatological data (LCD) observed from nearly 2,400 locations within the United States. The land-based, or surface, observations include air temperature, dew point, relative humidity, precipitation, wind speed and direction, visibility, atmospheric pressure, and types of weather occurrences [52]. Previous research shows broken rails are more likely to occur in cold weather due to tensile thermal stresses [1,42]. Thus, in this study, the minimum ambient temperature of the last 30 days is used as one input variable, in order to consider the influence of rail temperature. The temperature data from the nearest weather stations are allocated to segments of the railroad according to the geological information of the central point of segments.

4.2. *Feature extraction*

Feature extraction refers to the translation of raw data into the input variables (features). The following variables (Fig. 5) are generated as

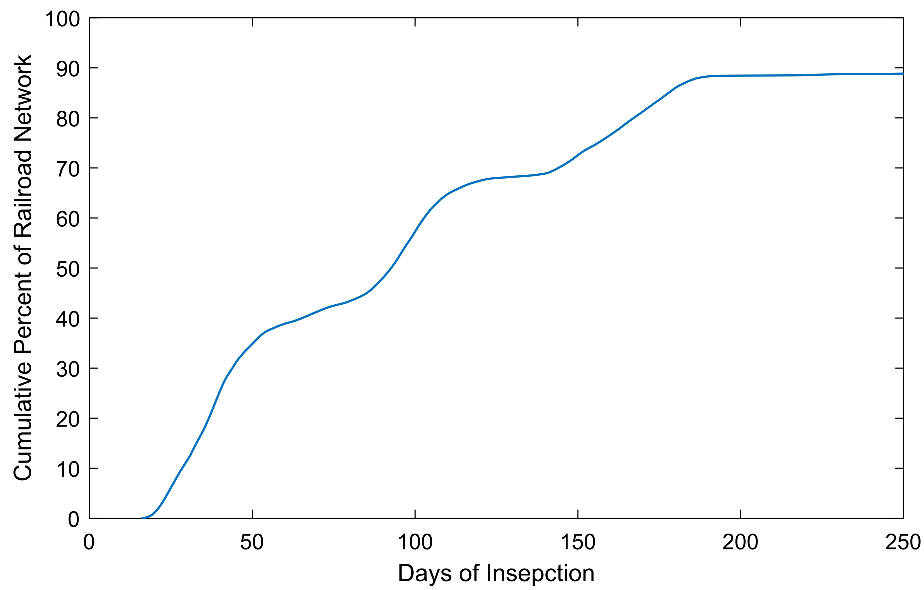

Fig. 4. Cumulative Distribution of Railroad Network with Ultrasonic Rail Flaw Inspection.

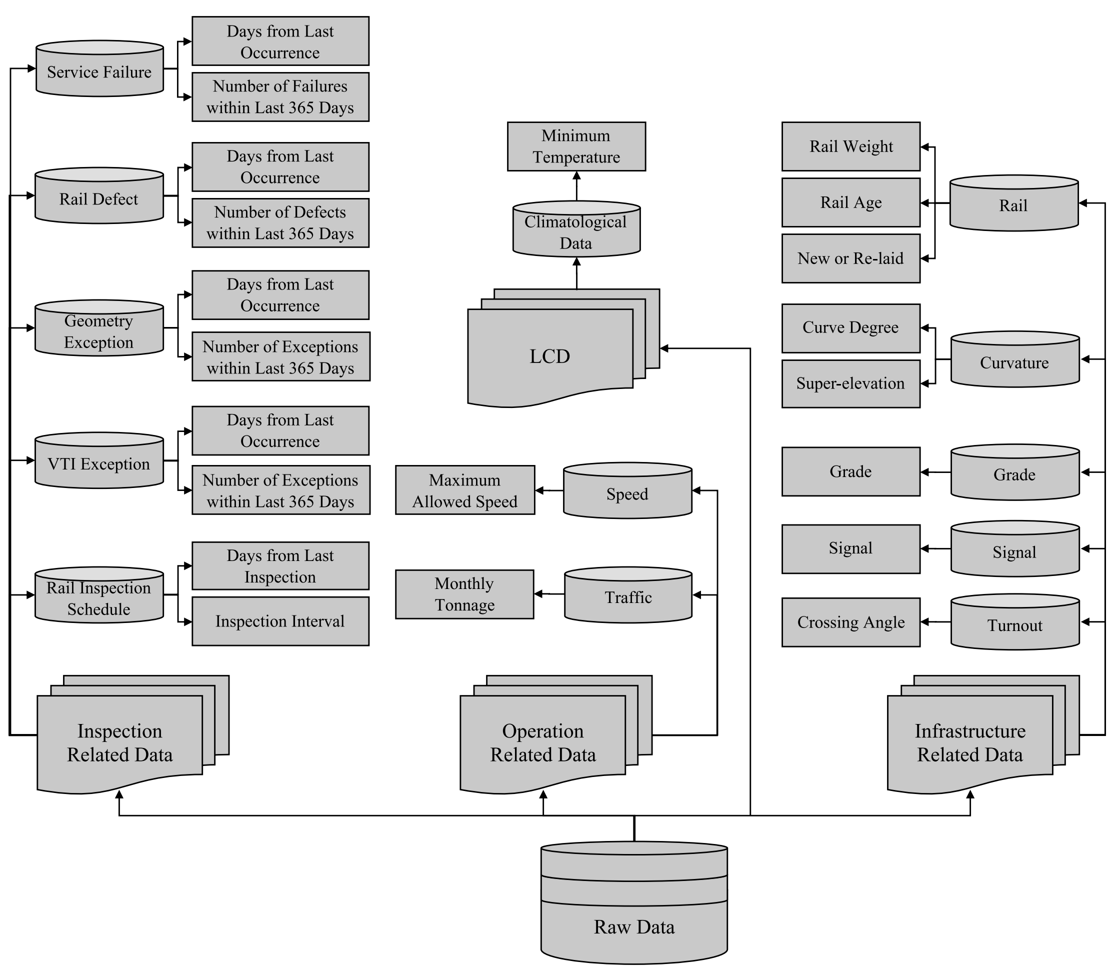

Fig. 5. Features Extracted from Raw Railroad Data and Climatological Data.

preliminary input variables from the raw data based on the literature review (Table 1). When it comes to inspection related information, they are separately defined as input features of the machine learning model. For instance, features associated with rail defects (excluding broken rail) are solely extracted from the rail defect database. While broken rail (rail service failure) is the output variable (prediction target) of the model, the past information of the broken rail is defined as input variables, such as the number of days from the last failure occurrence and the number of

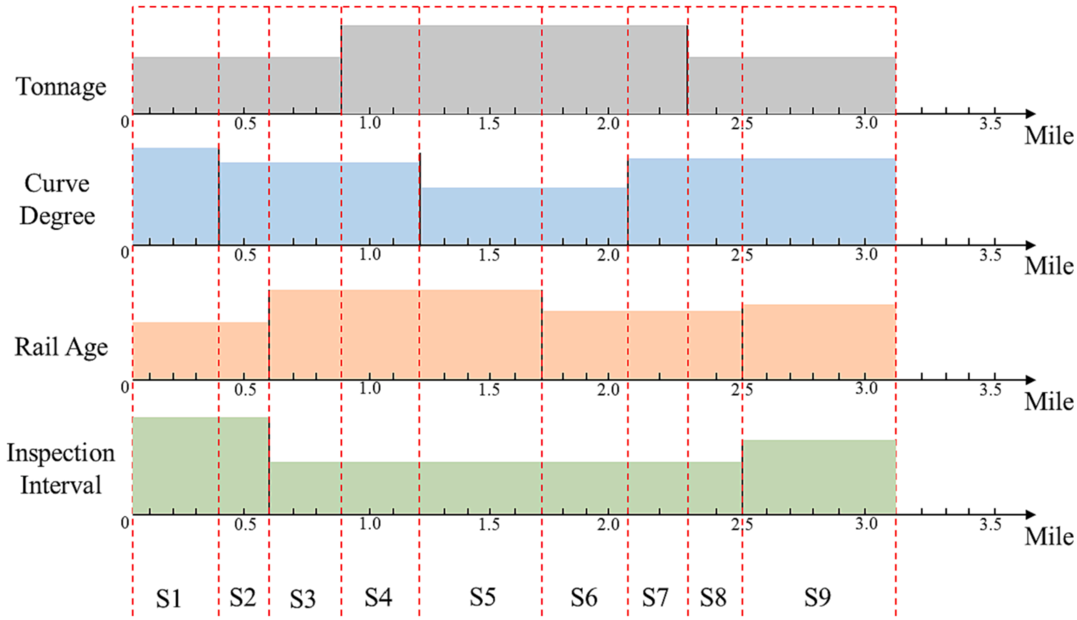

Fig. 6. A Hypothetical Illustrative Example of Feature-based Track Segmentation.

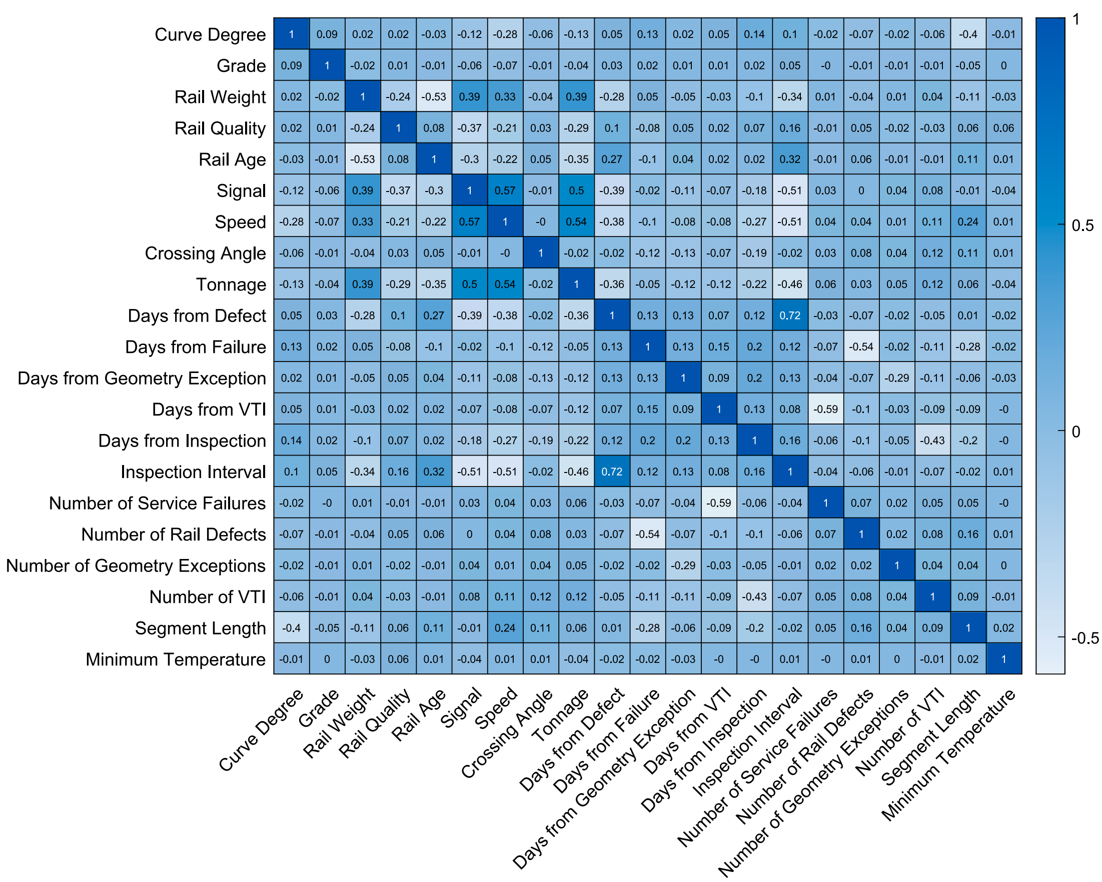

Fig. 7. Correlation Plot for the Majority of Input Variables.

rail failures within the last 365 days. Further, some combinations of attributes are added as additional features in the final integrated database. These include dynamic wheel load, product of rail age and degree of curvature, product of annual traffic density (in MGT) and rail weight, and quotient of annual traffic density and rail weight [10].

4.3. Track segmentation

Track segmentation is a process of delineating the railroad network into various segments. This section introduces track segmentation using the integrated dataset generated based on temporal and spatial information. The integrated data set is first divided into segments with a fixed

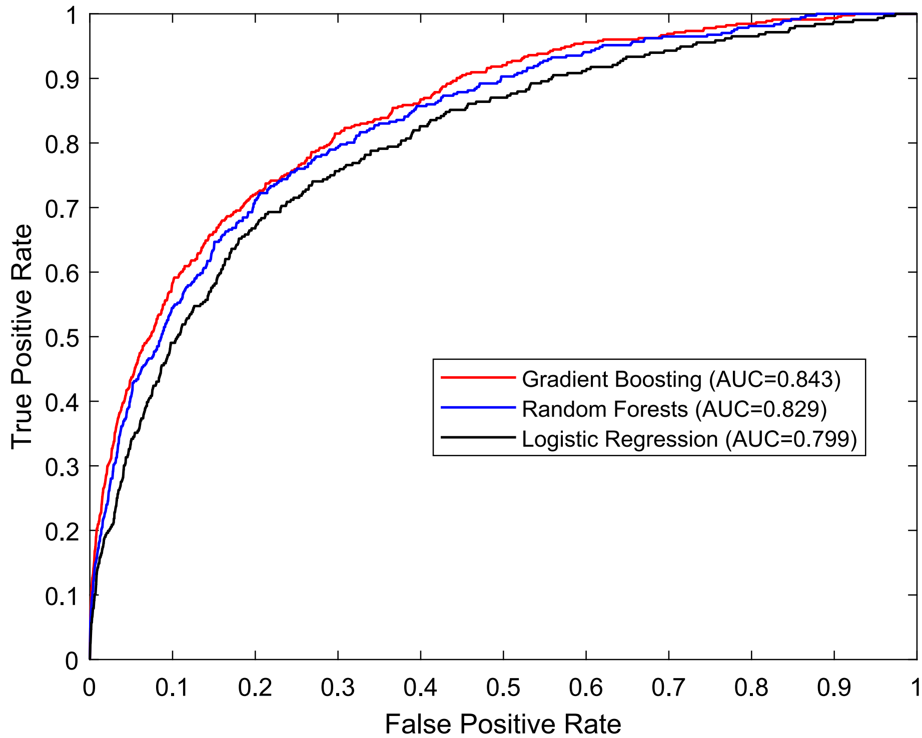

Fig. 8. Testing Performance of Proposed Models without Resampling.

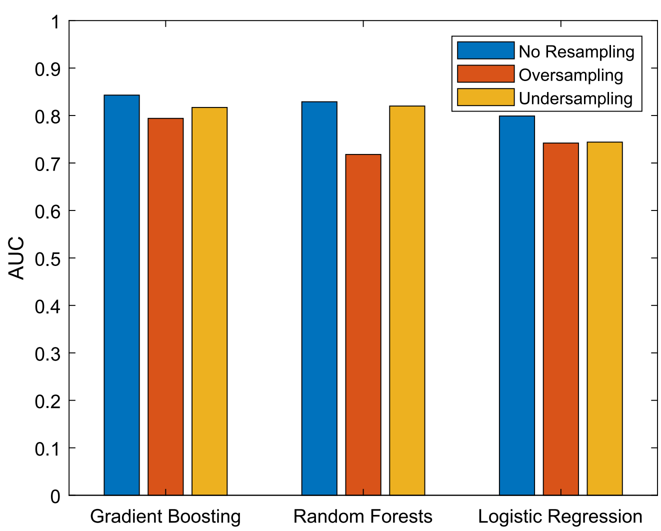

Fig. 9. AUC of Models under Different Scenarios.

length, which is defined as fixed-length track segmentation. The resolution of traffic data is 0.1 miles. Thus, in this paper, the primary segment length is set to 0.1 miles. While short segment length incurs high computation costs due to large records, long segment length results in an increase in the detection workload. Thus, in this paper, feature-based track segmentation is introduced to combine 0.1-mile segments with the same or similar attributes into longer segments, in order to reduce variances in the new data records and save computation costs. Fig. 6 demonstrates an example of feature-based track segmentation based on partial features selected. The track segmentation was performed according to the following steps:

Step 1: Generate 0.1-mile segments for the whole network. Each segment representing one record has a fixed length and can be uniquely identified by location information, including Prefix, Track Type, and Milepost.

Step 2: According to the location information of the segments, for each segment, map all of the time-independent attributes, including all of the infrastructure-related attributes and maximum allowable speeds.

Step 3: Timestamps that indicate the year and month of each record are introduced to specify the time attributes of the integrated data.

Step 4: According to the time information and geographic information, for each segment, map the time-dependent variables, including inspection data, traffic data, and climatological data.

Step 5: If adjacent segments have similar values of infrastructure attributes and operation attributes, they are combined into a longer segment. The maximum length of segments in this research is 0.5 miles.

4.4. *Correlation of input variables*

Pearson correlation analysis was conducted (Fig. 7). The results illustrate that most features can be directly used due to their low values of correlation with each other. The correlation coefficient between the days from last rail inspection and rail inspection interval is 0.72, indicating a strong linear relationship. This is reasonable since segments with greater inspection intervals are more likely to have more days since the last rail inspection. Of these two variables, only the inspection interval is selected as an input variable. Three attributes (speed, tonnage, and signal) are also related, since the territory with signal control tends to have higher speeds, which further contributes to larger traffic volumes. Finally, except for days from last rail inspection having a strong correlation with inspection interval, all the other single features and combinations of features (see details in section 4.2) are used as input variables in the model.

4.5. *Model results*

Using the aforementioned parameters as inputs, we develop three different machine learning methods that are trained based on data from 2013 to 2018. The model is used to predict the monthly broken rail occurrence in 2019. A comparison is made among these methods

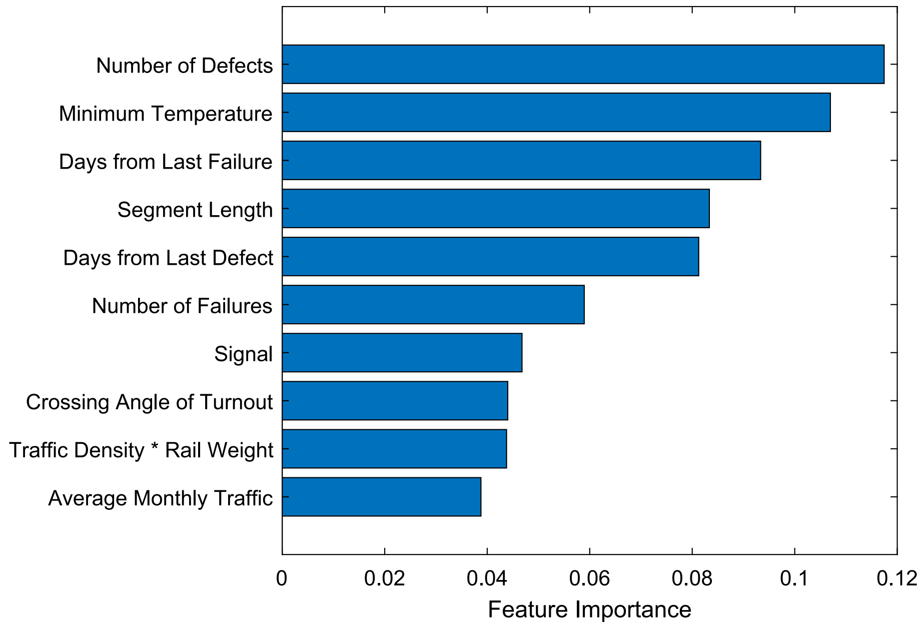

Fig. 10. Top Influencing Variables for Broken Rail Prediction.

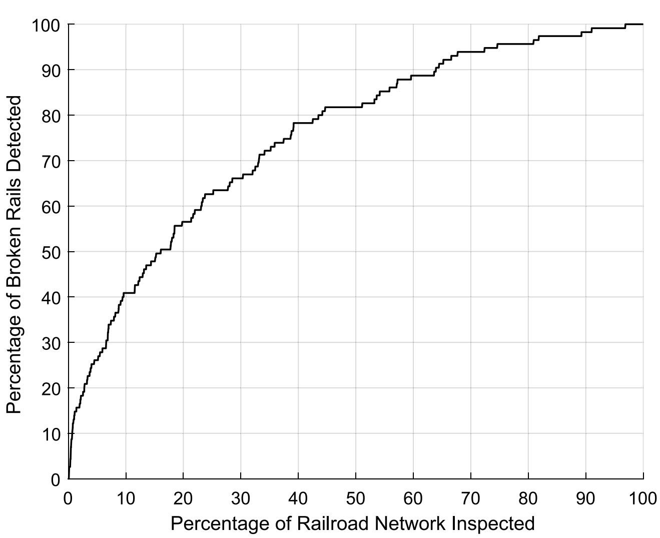

Fig. 11. Application Performance of Proposed Model for January 2019.

without data resampling in terms of the ROC and the AUC, as shown in Fig. 8. The gradient boosting model with a 0.843 AUC value outperforms Random Forests and Logistic Regression, whose values are 0.829 and 0.799, respectively. The training, validation, and testing AUC of the gradient boosting model are 0.876, 0.854, and 0.843, respectively.

Additionally, we apply the same sequence of methodologies using fixed-length track segmentation. In this case, all segments have a fixed length of 0.1 miles. The results show that although higher computation costs are needed to fit machine learning models using the integrated dataset with 0.1-mile segmentation, gradient boosting still demonstrates better performance than the other two algorithms. However, the AUC value of the testing dataset with fixed-length segmentation dropped from 0.843 to 0.828. Therefore, the feature-based track segmentation proposed in this research can effectively improve the performance of the proposed model.

Further analysis is conducted to study the effectiveness of extra variables, including minimum temperature and crossing angles of turnouts, which have not been considered in previous studies. The results show that model performance significantly decreases when minimum temperature and crossing angles of turnouts are excluded from the input variables of the machine learning models. While the gradient boosting still outperforms the other two algorithms, the AUC drops from 0.843 to 0.798, demonstrating the significance of extra variables in this research.

It is interesting to see that the performance of our gradient boosting model cannot be significantly improved with the implementation of data resampling (Fig. 9), while previous research [2] showed the benefits of resampling for model improvement. This might be due to the use of different datasets for different locations. Our model is for the entire network, while Ghofrani et al. [2] focused on high-density corridors.

Fig. 10 includes the top ten influencing variables for broken rail prediction. It indicates that the number of rail defects within the last 365 days, minimum ambient temperature for the last 30 days, days from last broken rail, segment length, and days from last rail defect are among the variables that significantly influence broken rail prediction results.

Temperature variations affect the tensional state of the rails. Broken rails are more likely to occur in the cold season, given all else is equal [2,43]. Among the top ten variables, two variables are associated with traffic density: the average monthly traffic and the product of annual traffic density and rail weight. Traffic density is related to wheel-rail interaction, which can affect rail degradation [44,45].

4.6. Model application

In the model application, the developed machine learning model can be used to identify the track segments that are most prone to broken rail occurrence. First, raw data are collected for generating input variables for the machine learning model. Then, the probability of broken rail occurrence for each segment is predicted using the developed model. Finally, segments with higher predicted risks of broken rail should be given priority during inspection and/or maintenance. The relationship between the percentage of the network scanned and the percentage of broken rails identified shows the performance of the model application. Fig. 11 illustrates an example of the model application for predicting broken rail occurrence in January 2019. When 10% of the railroad network is scanned using our proposed model, around 41% of the total number of broken rails can be found in those locations. Fig. 12 shows the predicted risk of broken rail over the partial network on the studied

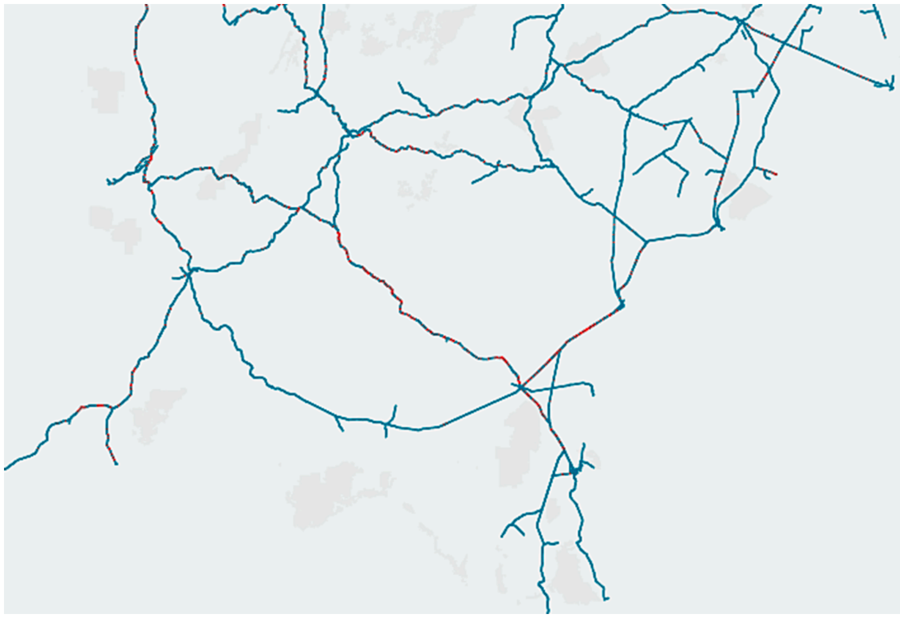

Fig. 12. Predicted Top 10% Risk (Red Segments) of Broken Rails over Part of Studied Railroad Network.

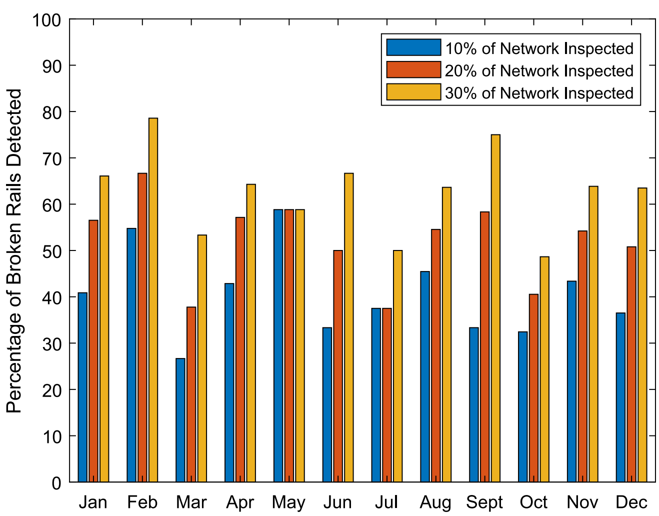

Fig. 13. Application Performance of Proposed Model in Different Months in 2019.

railroad in January 2019, where the segments with relatively higher probabilities of broken rail are colored in red.

Finally, the proposed model is applied to predict broken rails in different months for further evaluation of its generalization performance (Fig. 13). The model's performance varies in different months, which may be caused by the different numbers of broken rails in various months. However, it shows that the proposed model can be helpful in improving the efficiency of railroad inspection and maintenance in all months. The performance of the model in all months ranges from 26.7% to 58.8%, and the average performance is 40.4% when 10% of the railroad network is inspected. Between 37.5% and 66.7% of broken rails can be predicted, and the average performance is 51.9% when 20% of the railroad network is scanned. If 30% of the railroad network is inspected based on the proposed model, between 50.0% and 78.5% of broken rails would be identified, and the average performance in all months would be 62.7%.

5. Conclusions

Predicting the occurrence of broken rails has safety and economic benefits in the form of reducing accidents and service disruptions. The purpose of this study is to develop a data-driven prediction model with good performance for predicting broken rails one month in advance. With infrastructure data, operation data, inspection data, and weather condition data for the railroad from 2013 to 2019 being considered in the machine learning models, gradient boosting outperformed the other two methods for our datasets. To avoid data leakage and resultant over-optimistic estimations of the model, the testing data used for evaluating the performance of the proposed model was generated by partitioning all of the data according to time sequences. Featured-based track segmentation is proven to be an effective way of saving computing costs and improving the performance of the proposed model.

Compared to previous research, extra variables including minimum temperature and crossing angles of turnouts are extracted and considered as input variables to further improve the model's prediction performance. Regarding feature importance, this study identifies that the number of detected rail defects within the last 365 days, minimum ambient temperature of the last 30 days, days from last broken rail, segment length, traffic density and other factors have significant influences on prediction results. Using the model proposed in this paper, on average, 40.4% of broken rails can be predicted one month in advance when focusing on 10% of the railroad network scanned. The model can potentially be used to prioritize inspection and maintenance activities on broken-rail-prone locations, and thus further improve infrastructure safety given budgetary constraints.

6. Future work

This section discusses the future directions of this research. First, more variables can be included to further improve the performance of the prediction model. As broken rails are subject to various factors, additional variables such as substructure conditions could be incorporated for future model improvement. For example, track segments in the highly fouled territory (greater ballast fouling index) are more likely to have greater dynamic loads on the rail, leading to a higher risk of broken rails.

A more fine-tuned model may perform better in broken rail prediction than the current model, using the same methodology proposed in this research. The strategy of track segmentation, for example, can be further studied. The influence of the diversity of involved features for segmentation needs to be further analyzed to select the best appropriate features for delineating the railroad network. Another alternative approach for fine-tuned segmentation is associated with segment length such as setting appropriate upper and lower thresholds of segment length. Additional research might be necessary to better handle the imbalanced data mining issues due to the scarcity of broken rail occurrences. For instance, a weighted evaluation function that can consider the weight of segments with occurrence of broken rails (minority class) can be used in the machine learning model.

Last but not least, data-driven optimization of inspection and maintenance planning requires further exploration to provide a detailed inspection and spot maintenance strategy for railroad staff.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded through a contract by the Federal Railroad Administration (693JJ620C000017). However, the authors are solely responsible for all views and analyses in this research.

References

[1] Liu, X., Dick, C. T., Lovett, A., Saat, M. R., & Barkan, C. P. (2013, October). Seasonal effect on the optimization of rail defect inspection frequency. In *Rail Transportation Division Conference* (Vol. 56116, p. V001T01A008). American Society of Mechanical Engineers.

[2] F. Ghofrani, H. Sun, Q. He, Analyzing Risk of Service Failures in Heavy Haul Rail Lines: A Hybrid Approach for Imbalanced Data, Risk Analysis. (2021).

[3] X. Liu, Statistical causal analysis of freight-train derailments in the United States, Journal of transportation engineering, Part A: Systems 143 (2) (2017) 04016007.

[4] Schafer, D. H., & Barkan, C. P. (2008, September). A prediction model for broken rails and an analysis of their economic impact. In *Proc., American Railway Engineering and Maintenance of Way Association (AREMA) Annual Conference*.

[5] Milo, D., Principe, L., Deng, J., Zhou, K., & Liu, X. (2018, April). A Literature Review of Rail Defect Causal Factors. In *ASME/IEEE Joint Rail Conference* (Vol. 50978, p. V001T06A005). American Society of Mechanical Engineers.

[6] C. Kang, S. Schneider, M. Wenner, S. Marx, Experimental investigation on the fatigue behaviour of rails in the transverse direction, Construction and Building Materials 272 (2021) 121666.

[7] Surface Transportation Board. (n.d.). Retrieved October 23, 2021, from https://www.stb.gov/reports-data/economic-data/.

[8] C.T. Dick, Factors affecting the frequency and location of broken railway rails and broken rail derailments, University of Illinois at Urbana-Champaign), 2001. Doctoral dissertation.

[9] Dick, C. T., Barkan, C. P., Chapman, E., & Stehly, M. P. (2002). Predicting the occurrence of broken rails: a quantitative approach. In Proceedings of the 2002 Annual Conference, American Railway Engineering and Maintenance of Way Association (AREMA), Washington DC.

[10] C.T. Dick, C.P. Barkan, E.R. Chapman, M.P. Stehly, Multivariate statistical model for predicting occurrence and location of broken rails, Transportation Research Record 1825 (1) (2003) 48–55.

[11] Schafer, D. H., & Barkan, C. P. L. (2008, May). A hybrid logistic regression/neural network model for the prediction of broken rails. In *Proceedings of the 8th world congress on railway research, Seoul, Korea*.

[12] S. Vesković, J. Tepić, M. Ivić, G. Stojić, S. Milinković, Model for predicting the frequency of broken rails, Metalurgija 51 (2) (2012) 221–224.

[13] G. Chattopadhyay, S. Kumar, Parameter estimation for rail degradation model, International Journal of Performability Engineering 5 (2) (2009) 119.

[14] A. Jamshidi, S.F. Roohi, A. Núñez, R. Babuska, B. De Schutter, R. Dollevoet, Z. Li, Probabilistic defect-based risk assessment approach for rail failures in railway infrastructure, IFAC-PapersOnLine 49 (3) (2016) 73–77.

[15] A. Jamshidi, S. Faghih-Roohi, S. Hajizadeh, A. Núñez, R. Babuska, R. Dollevoet, Z. Li, B. De Schutter, A big data analysis approach for rail failure risk assessment, Risk analysis 37 (8) (2017) 1495–1507.

[16] L. Bai, R. Liu, F. Wang, Q. Sun, F. Wang, Estimating railway rail service life: A rail-grid-based approach, Transportation Research Part A: Policy and Practice 105 (2017) 54–65.

[17] F. Ghofrani, N.K. Chava, Q. He, Forecasting risk of service failures between successive rail inspections: a data-driven approach, Journal of Big Data Analytics in Transportation 2 (1) (2020) 17–31.

[18] H. Li, D. Parikh, Q. He, B. Qian, Z. Li, D. Fang, A. Hampapur, Improving rail network velocity: A machine learning approach to predictive maintenance, Transportation Research Part C: Emerging Technologies 45 (2014) 17–26.

[19] Lovett, A. H., Dick, C. T., Ruppert Jr, C., Saat, M. R., & Barkan, C. (2013, April). Development of an integrated model for the evaluation and planning of railroad track maintenance. In *ASME/IEEE Joint Rail Conference* (Vol. 55300, p. V001T01A003). American Society of Mechanical Engineers.

[20] Kubat, M., & Matwin, S. (1997, July). Addressing the curse of imbalanced training sets: one-sided selection. *In Proceedings of the Fourteenth International Conference on Machine Learning* (Vol. 97, pp. 179-186).

[21] Japkowicz, N. (2000, June). The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence* (Vol. 56).

[22] Spelmen, V. S., & Porkodi, R. (2018, March). A review on handling imbalanced data. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-11). IEEE.

[23] R.A. Sowah, M.A. Agebure, G.A. Mills, K.M. Koumadi, S.Y. Fiawoo, New cluster undersampling technique for class imbalance learning, International Journal of Machine Learning and Computing 6 (3) (2016) 205–214.

[24] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[25] Wong, S. C., Gatt, A., Stamatescu, V., & McDonnell, M. D. (2016, November). Understanding data augmentation for classification: when to warp?. In *2016 international conference on digital image computing: techniques and applications (DICTA)* (pp. 1-6). IEEE.

[26] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE.

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

[28] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[29] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[30] Wang, Y., & Ni, X. S. (2019). A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433*.

[31] Branco, P., Torgo, L., & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*.

[32] Jabbar, H., & Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, 163-172.

[33] Rocks, J. W., & Mehta, P. (2020). Memorizing without overfitting: Bias, variance, and interpolation in over-parameterized models. *arXiv preprint arXiv:2010.13933*.

[34] E. Briscoe, J. Feldman, Conceptual complexity and the bias/variance tradeoff, Cognition 118 (1) (2011) 2–16.

[35] R. Pautrat, K. Chatzilygeroudis, J.B. Mouret, in: May). Bayesian optimization with automatic prior selection for data-efficient direct policy search, IEEE, 2018, pp. 7571–7578.

[36] R. Chen, J. Chen, P. Wang, J. Fang, J. Xu, Impact of wheel profile evolution on wheel-rail dynamic interaction and surface initiated rolling contact fatigue in turnouts, Wear 438-439 (2019) 203109.

[37] A.M. Zarembski, D. Einbinder, N. Attoh-Okine, Using multiple adaptive regression to address the impact of track geometry on development of rail defects, Construction and Building Materials 127 (2016) 546–555.

[38] R. Mohammadi, Q. He, F. Ghofrani, A. Pathak, A. Aref, Exploring the impact of foot-by-foot track geometry on the occurrence of rail defects, Transportation research part C: emerging technologies 102 (2019) 153–172.

[39] D. Li, E.T. Selig, Wheel/track dynamic interaction: track substructure perspective, Vehicle System Dynamics 24 (sup1) (1995) 183–196.

[40] D.P. Connolly, G. Kouroussis, O. Laghrouche, C.L. Ho, M.C. Forde, Benchmarking railway vibrations–Track, vehicle, ground and building effects, Construction and Building Materials 92 (2015) 64–81.

[41] J. Xiao, D. Zhang, K. Wei, Z. Luo, Shakedown behaviors of railway ballast under cyclic loading, Construction and building materials 155 (2017) 1206–1214.

[42] Z. Zhang, X. Liu, H. Hu, Statistical analysis of seasonal effect on freight train derailments, Journal of Transportation Engineering, Part A: Systems 147 (10) (2021) 04021073.

[43] B. Suárez, P. Rodriguez, M. Vázquez, I. Fernández, Safety assessment of underground vehicles passing over highly resilient curved tracks in the presence of a broken rail, Vehicle system dynamics 50 (1) (2012) 59–78.

[44] J.W. Ringsberg, Life prediction of rolling contact fatigue crack initiation, International Journal of fatigue 23 (7) (2001) 575–586.

[45] R.D. Fröhling, Wheel/rail interface management in heavy haul railway operations—applying science and technology, Vehicle System Dynamics 45 (7–8) (2007) 649–677.

[46] F. Ghofrani, A. Pathak, R. Mohammadi, A. Aref, Q. He, Predicting rail defect frequency: An integrated approach using fatigue modeling and data analytics, Computer-Aided Civil and Infrastructure Engineering 35 (2) (2020) 101–115.

[47] U. Zerbst, M. Schödel, R. Heyder, Damage tolerance investigations on rails, Engineering Fracture Mechanics 76 (17) (2009) 2637–2653.

[48] X. Ge, L. Ling, X. Yuan, K. Wang, Effect of distributed support of rail pad on vertical vehicle-track interactions, Construction and Building Materials 262 (2020) 120607.

[49] FRA (Federal Railroad Administration). (2015). Track inspector rail defect reference manual.
[50] X. Liu, M.R. Saat, C.P. Barkan, Analysis of causes of major train derailment and their effect on accident rates, Transportation Research Record 2289 (1) (2012) 154–163.
[51] Clark, D., Toth, T., Dick, M., & Maldonado, R. (2015, October). V/TI Monitor Cluster Analysis and Implementation. In *AREMA 2015 annual conference & exposition, Minneapolis (USA)*.
[52] National Centers for Environmental Information (NCEI). (2021, October 13). *U.S. Local Climatological Data (LCD)*. U.S. Local Climatological Data (LCD). Retrieved April 1, 2022, from https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00684.