Received May 8, 2020, accepted May 22, 2020, date of publication June 2, 2020, date of current version June 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2999385

Rail Weld Defect Prediction and Related Condition-Based Maintenance

NAN YAO[1], YUEJUN JIA[2], AND KAI TAO[1]
[1]China Academy of Railway Sciences Corporation Ltd., Beijing 100081, China
[2]Ningbo Track Maintenance Division, China Railway Shanghai Group Company Ltd., Ningbo 315000, China

Corresponding author: Nan Yao (yaonan0618@163.com)

This work was supported in part by the Ministry of Science and Technology of China, in part by the Infrastucture Inspection Institute, China Academy of Railway Sciences Corporation Ltd., and in part by the Comprehensive Inspection and Monitoring on Rail Defects Project under Project 2016YFF0103705.

ABSTRACT Rail weld defects are major threats to railroad transportation. Enormous resources have been required for related maintenance. This paper presents a creative solution to predict weld defects and to classify railroads into different conditions based on the predictions. The results are based on features extracted from manufacturing technologies of welds, from related materials and from influential factors in the environments. Features such as marks for welding engineers are defined. Maintenance can be selectively implemented based on the predicted conditions. Safety is the foundation of the railroad business, and a very strict safety requirement is utilized as one of the main constraints in this research. Additionally, 11 key risk factors leading to rail defects and their risk levels are identified. Extreme learning machine (ELM), random forest, logistic regression, principal component analysis (PCA), support vector machine (SVM) and other data science approaches are utilized. The evaluation results show that the related rail maintenance workload can decrease significantly under high safety standards. Labor costs of weld inspection will be reduced substantially because of the decreased workload for the sections predicted to not have any defects with a 100% recall rate (approximately 30% of the total sections), contributing to a massive cost reduction. Consequently, rail companies are expected to achieve enhanced management and operation.

INDEX TERMS Condition-based maintenance, extreme learning machine, logistic regression, rail weld defect prediction, random forests, support vector machine.

I. INTRODUCTION

Rail defect research is pivotal for railway companies [1]. Therefore, they have put plenty of effort into rail defect detection and related maintenance [2]–[11]. This research presents a new type of data-driven method for rail defects. It entails the prediction of rail defects and related implications for railroad management.

Currently, time-based maintenance is widely used in the railroad industry. However, this type of work causes tremendous waste because it requires a heavy maintenance workload at the same level for each section of a railroad line (a railroad line can be divided into multiple sections), but it is commonly accepted that some sections of a railroad line could be significantly better or worse than the others. New research has also shown that predictive maintenance is the most promising maintenance strategy for railroads [12]–[18]. At the end of 2019, China had more than 139,000 kilometers of railroads. The tracks are connected by welding, and at least one weld is needed for every 25 meters to 100 meters of track. In China, it is estimated that more than 120,000 labor days per month (equal to hiring 4,000 workers to work 30 days) are needed to finish related weld inspection work. Thus, a quantitative analysis to classify railroad sections by weld conditions and to implement predictive inspection/maintenance is desirable.

Track is one of the most critical components for railroads, and track defects may lead to severe issues, including derailments. According to our calculations, approximately 52.6% of rail defects occurred on welding joints, which are considered the weakest parts of tracks [19]–[24]. However, time-based maintenance is widely applied in the related work. The work is scheduled based on highly conservative estimates for all sections of railroad lines. If we can reallocate resources based on the predicted conditions of the sections (e.g., divide sections into sections in better condition and sections in worse condition based on weld conditions), costs and work time may be saved significantly for the sections in better condition.

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

FIGURE 1. Rail weld defects.

However, such condition classification for welds before the start of inspection/maintenance has not been completed. Currently, there is no published research on predicting the conditions based on all the major features extracted from manufacturing technologies of welds, from materials and from influential factors in the environments. In addition, the mechanisms between the defects and their indicative factors are so complicated that people do not understand them clearly. Multiple data analysis methods have been applied to these areas [19]–[23]. However, these methods have limitations. One of the most recent studies presented a Pareto-based maintenance decision system using the Hilbert spectrum, but the result was mainly dependent on the dynamic response from axle box acceleration measurements [24]. The acceleration data on which their model was built correspond to vibrations caused by track irregularities, so they are not sufficient to show the internal conditions and production quality of welds and to predict the occurrence of weld defects [25]. Instead, we utilize features extracted from a wide range of easily accessible data and machine learning methods to solve weld problems. In another study, squat defects and ballast defects were treated using optimization methods and condition-based maintenance [12], [26]. Squats occur on the surface of tracks, and ballast is a substance below tracks. Thus, these two types of defects are different from weld defects. Another researcher also proposed a framework for rail surface defect prediction using machine learning algorithms. The research is limited to the surface defects, and it is not related to manufacturing technologies of welds and the materials of welds [27]. Other researchers have also not claimed any success or viable solutions to the problems we are working on [2]–[11], [21]–[22], [28]–[30]. We first extract all the key features related to the problems and first utilize data mining approaches for weld defect prediction problems. In addition, manufacturing technologies of welds have been creatively analyzed, and 11 key risk factors leading to rail defects and their risk levels are identified. Using the predictive models presented in this research, railroad maintenance can be decreased significantly under very high safety standards. In addition, during special periods such as the Covid-19 period, the number of engineers inspecting welds may have to decrease to satisfy health concerns.

All the prediction results are under 100% recall rate which means an extremely low probability for defect occurrence. In addition, according to our newest database, more than 95% of the defects are minor defects that do not require any repair. Therefore, the risk for the rail sections predicted to not have any defects is very low. Traditionally, inspection workloads for all the sections are equally heavy because time-based maintenance. Compare to this, the inspection workload will be reduced significantly for the low-risk sections based on the models proposed in the research. The workload decrease for these sections is around 50% at the Ningbo maintenance department involved in this research. The low-risk sections account for approximately 29.82%-31.58% of the total sections. This suggests a massive cost reduction for the intense weld inspection work (more than 120,000 labor days per month in China) conducted in the railroad world.

Our related research was presented at the World Congress on Railway Research in Tokyo, Japan [31]. However, the work submitted was only an analysis focusing on the correlation between track geometry and rail defects. The modeling and evaluation in the current paper are also significantly different.

II. METHODS

The main output is the predicted condition (a better condition or a worse condition) of a rail section. The inputs are introduced in Table 1. The recall rate should be high enough to satisfy safety requirements. Logistic regression, random forests, extreme learning machine (ELM) and support vector machine (SVM) are the main modeling methods.

A. DATA ACQUISITION AND PREPARATION

Data are collected from railroad companies that manage regular-speed railroads and high-speed railroads. A relational database is developed to manage the data. According to our business understanding and experience from railroad experts, each variable is defined below. Based on the definitions, the data are processed.

The target variable is defects, and the remaining variables are the predictors.

The rows with missing values are deleted. The descriptive statistics are also determined, and all the predictors are normalized. A total of 974 rail sections are included in the training dataset, including validation data. Data from all the related major lines were recently updated significantly, and they are used as the testing dataset. All the data are processed by using R software version 3.6.1.

B. COLLINEARITY

Collinearity is a problem caused by correlations among predictors. The correlations may lead to inaccurate models developed by these predictors.

Therefore, the collinearity may influence the modeling. The logistic regression, SVM and ELM models cannot eliminate this influence directly [35]. For logistic regression, backward stepwise selection is used to solve the problem. For the SVM and ELM models, principal component analysis is utilized to eliminate the collinearity. The selected principal

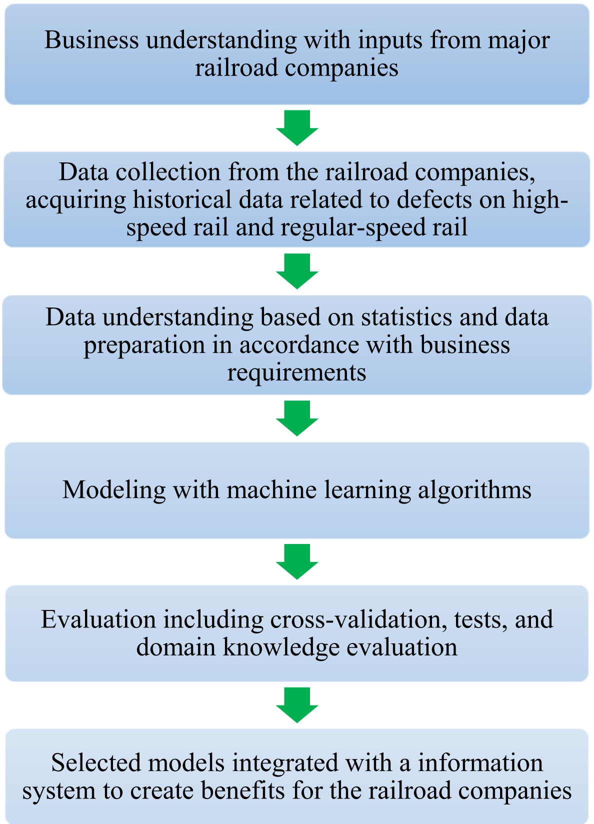

FIGURE 2. Overall view for proposed approach.

components serve as new variables. Along with the predictors that are not in the principal components, they will be the inputs for the SVM and ELM models. However, predictions made by the random forest model are not sensitive to the collinearity [36].

C. LOGISTIC REGRESSION

Logistic regression is a generalized linear regression model. It is easy to apply and explain. In logistic regression, the target variable is a probability: how likely a successful prediction is to occur. The relationship between the target variable and the predictors is shown in formula [37]:

$$q = \Pr(y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)}. \quad (1)$$

In this formula, $q(0 \leq q \leq 1)$ is the target variable. $X_1, X_2, \cdots, X_p$ are the predictors. By employing the maximum likelihood estimation, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ can be decided. However, it is necessary to analyze the impacts caused by multiple collinearity.

AIC values are calculated to evaluate the logistic-regression model before/after the backward stepwise selection. Moreover, for the best logistic-regression model, the importance of each predictors can be evaluated exactly. The detailed evaluation will be presented in the Results and Evaluation part.

TABLE 1. Definitions of variables [31]–[34].

Variable	Terminology	Explanation in this Research
Direction	Direction of Traffic	Up track: 1; down track: 2; single track: 3.
Start_mile	Starting Point of a Rail Section	Location information can be represented to show geographical features and other general information in the section.
End_mile	Ending Point of a Rail Section	
Pave_date	Date of Pavement	The exact month when the track in the section was paved
L_temp	Stress-free Temperature of Left Track	Average stress-free temperature of left track in centigrade in the section.
R_temp	Stress-free Temperature of Right Track	Average stress-free temperature of right track in centigrade in the section.
Load	Gross Tonnage per Kilometer	For the section, accumulated gross tonnage per kilometer in million tons.
TQI	Track Quality Index	Track Quality Index is the sum of standard deviations of seven items: irregularity of longitudinal level of left track, irregularity of longitudinal level of right track, irregularity of the cross level, irregularity of alignment of left track, irregularity of alignment of right track, track gauge, track twist. Calculated Track Quality Index for each rail section. Then, calculate the average value of the Track Quality Index.
Curve_number	Number of Curves	The number of curves in the section.
Curve_radius	Weighted Radius of Curves	In the section, the sum of the length of each curve divided by its radius.
Slope_number	Number of Slopes	The number of slopes in the section.
Slope_value	Weighted Value of Grade of Slopes	In the section, calculate the sum of the length of a slope times the grade of the slope; then, divide that sum by the total length of the slopes.
Grind_number	Number of Grinding	The number of grinding times in the section
YDH_number	Number of Flash-butt Mobile Welding	The amount of flash-butt mobile welding in the section.
YDH_rain_number	Rain Number during Flash-butt Mobile Welding	The amount of rain in the section when flash-butt mobile welding.
YDH_temp	Track Temperature during Flash-butt Mobile Welding	The average track temperature during flash-butt mobile welding in the section, in centigrade.
YDH_DDL	Displacement in Upset of Flash-butt Mobile Welding	Average displacement in a certain step of flash-butt mobile welding in the section.
LRH_number	Number of Alumino Thermit Welding	The amount of alumino thermit welding in the section.
LRH_rain_number	Rain Number during Alumino Thermit Welding	The amount of rain in the section when alumino thermit welding.
LRH_date	The Date when Alumino Thermit Welding was completed	The earliest date that alumino thermit welding was completed in the section.
LRH_left_model	Materials in front of Alumino Thermit Welds	If the material type in the section belongs to U75V, the type is 1; otherwise, it is 0.

TABLE 1. ***(Continued.)*** **Definitions of variables [31]–[34].**

Variable	Terminology	Explanation in this Research
LRH_right_model	Materials behind Alumino Thermit Welds	If the material type in the section belongs to U75V, the type is 1; otherwise, it is 0.
LRH_weather	Weather during Alumino Thermit Welding	Rainy and snowy: 4; Snowy: 3; Rainy: 2; Other: 1
LRH_before_temp	Track Temperature before Alumino Thermit Welding	The minimal track temperature before alumino thermit welding in the section, in centigrade.
LRH_after_temp	Track Temperature after Alumino Thermit Welding	The minimal track temperature after alumino thermit welding in the section, in centigrade.
LRH_abs_temp	Track Temperature Difference during Alumino Thermit Welding	Average absolute value of (LRH_after_temp-LRH_before_temp) in the section, in centigrade
LRH_up_width	Width of Weld Top after Alumino Thermit Welding	Width of Weld Top after alumino thermit welding
LRH_low_width	Width of Weld Bottom after Alumino Thermit Welding	Width of Weld Bottom after alumino thermit welding
LRH_QGL	Camber Control for Alumino Thermit Welding	Camber Control level for alumino thermit welding
LRH_reaction_time	Reaction Time for Alumino Thermit Welding	Reaction time for alumino thermit welding
LRH_quiet_time	Quiet Time for Alumino Thermit Welding	Quiet time for alumino thermit welding
LRH_User_fraction	Marks for Engineers (implementing alumino thermit welding; the other type of welding is conducted by machines automatically)	Grading based on the Stars (performance evaluation results reflecting skill levels) and experience of the engineers. For a single weld, the mark=the mark of rail alignment engineer*0.3+the mark of molding and sanding engineer*0.3+the mark of pre-heating engineer*0.3+the mark of grinding engineer *0.1 The mark for a section=Total marks of the whole section/amount of welds in the section
Line_Type	Line Types	Regular-speed rail is 2, and high-speed rail is 1.
Speed_Grade	Speed Levels	The highest design speed of railroad lines
Defect	Rail Defect Status	If any defects occurred and are verified by railroad engineers in a section, it is coded as 1; if no defects are verified in the section, it is coded as 0.

Note: The locations and the lengths of the sections are defined based on maintenance requirements from rail companies.

D. RANDOM FOREST

Random forest [38] is a machine learning approach combining theories of bagging ensemble learning with random subspace methods [36], [39]. Thus, it may improve the learning system. Random forest is not sensitive to multiple collinearity. The results are robust to various types of datasets [40]. Random forest in this paper are formed as follows [41]:

Let N = the number of samples in the training dataset and M = the number of varieties in the training dataset.

1. Conduct sampling with replacement N times from the training dataset, forming a new training dataset. The unselected samples will be deployed in initial predictions, evaluating errors of the model;
2. For each node, select m features randomly. The selected features will lead to decisions on each node. The optimal split of the trees will be calculated based on the different features;
3. Each decision tree is fully developed without trimming;
4. Repeat the above steps to construct other decision trees until the number of required trees is reached. The number of decision trees is adjusted based on optimization goals;
5. Each decision tree is utilized as a basic classifier to process ensemble learning, generating an integrated classifier. Predictors are input into the model to be classified. The output is decided by voting in which each decision tree gives its vote on the classification.

To achieve business objectives, it is important to find the optimal number of trees and the most suitable quantity of nodes.

E. EXTREME LEARNING MACHINE NEURAL NETWORK

Extreme learning machine (ELM) is an easy-to-use but effective single-hidden-layer feedforward network (SLFN) algorithm [42]. ELM is also applied to find the relationship between the predictors and the categorical results.

Compared with traditional neural networks, ELM can provide a faster learning speed and a more favorable generalization. Formidable advantages have been manifested in various industries [35]. However, for the model, it is also necessary to consider the impacts caused by multiple collinearity [37].

The fundamental principles of ELM are described briefly [43]. This is an algorithm with three steps. For a given training dataset $TrainData = \{(x_i, t_i)|x_i \in R_n, t_i \in R_m, i = 1, 2, \cdots, N\}$, the hidden node output function is $G(a, b, x)$, and the number of hidden nodes is L. The three steps are summarized as follows:

Step 1: Assign values to the hidden node parameters randomly: (a_i, b_i), $i = 1, 2, \cdots, L$.

Step 2: Calculate the hidden layer output matrix, which is named H,

$$\sum\nolimits_{i=1}^{L} \beta_i G\left(a_i, b_i, x_j\right) = t_j, \quad j = 1, \cdots, N. \tag{2}$$

This is equal to $H\beta = T$. The ith column in H is the output from the ith hidden node, and the corresponding inputs are $x_1, x_2, \cdots, x_N$.

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} G(a_1, b_1, x_1) \cdots G(a_L, b_L, x_1) \\ \vdots \\ G(a_1, b_1, x_N) \cdots G(a_L, b_L, x_N) \end{bmatrix}_{N\times L};$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L\times m}, \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N\times m}. \tag{3}$$

Step 3: Calculate the output weights β:

$$\beta = H^{\dagger}T. \tag{4}$$

H† is the Moore-Penrose generalized inverse of hidden layer output matrix H.

Additionally, the number of nodes will be adjusted based on optimization objectives, which are defined based on business goals.

$G(a, b, x)$ is an activation function selected from the following:

TABLE 2. Alternatives for the activation function in ELM.

Function Type	Formula
Sigmoid	$G(a,b,x) = \dfrac{1}{1+exp(-(ax+b))}$
Sin	$\mathrm{G(a,b,x) = sin(ax+b)}$
Radial Basis	$G(a,b,x) = g(\lVert ax+b-c\rVert)$
Hard Limit	$G(a,b,x) = \begin{cases} 0, & ax+b<0 \\ 1, & ax+b\geq 0 \end{cases}$
Symmetric Hard Limit	$G(a,b,x) = \begin{cases} -1, & ax+b<0 \\ 1\ , & ax+b\geq 0 \end{cases}$
Symmetric Saturating Linear	$G(a,b,x) = \begin{cases} 0\ , & ax+b<-1 \\ ax+b, & -1\leq ax+b\leq 1 \\ 1\ , & ax+b\geq 1 \end{cases}$
Hyperbolic Tangent Sigmoid	$G(a,b,x) = \dfrac{\exp(ax+b)-\exp(-(ax+b))}{\exp(ax+b)+\exp(-(ax+b))}$
Triangular Basis	$G(a,b,x) = \begin{cases} 1-abs(ax+b), -1\leq ax+b\leq 1 \\ 0\ , \quad otherwise \end{cases}$
Rectifier Linear Unit	$G(a,b,x) = \max(0, ax+b)$
Linear	$G(a,b,x) = ax+b$
Gauss	$G(a,b,x) = exp(-b\lVert x-a\rVert)$

F. SUPPORT VECTOR MACHINE [44]

Support vector machine (SVM) is another machine learning approach. It is applied to classification tasks. Using SVM, the optimal hyperplane dividing two categories can be found by maximizing the distance between the closest points in both categories.

If there is a hyperplane that linearly separates samples, then define that x_i is a vector and that $y_i = 1\ or\ -1$ serves as a classification mark. Then, the optimal hyperplane represented as $\boldsymbol{w} * \boldsymbol{x} + b = 0$ can be found. Therefore, SVM solves the following programming problem:

$$\begin{aligned} & min\ \frac{1}{2}\lVert w\rVert^2 + C\sum_{i=1}^{l}\xi_i \\ & s.t.\ y_i\left(w\cdot x_i+b\right) \geq 1-\xi_i, \quad i=1,\cdots,l, \\ & \quad \xi_i \geq 0, i=1,\cdots,l. \end{aligned} \tag{5}$$

$w = \sum_{i=1}^{n}\alpha_i y_i x_i$ is a linear combination of all the support vectors. $\alpha_i\,(i=1,\cdots,n)$ is a Lagrange multiplier, and C is a penalty term. $\xi_i(i=1,\cdots,l)$ is a relaxation variable, and b is a constant. If there are no hyperplanes that linearly separate samples, generally the samples will be mapped to a higher-dimensional space by a kernel function. In this space, the samples can be linearly separated effectively. Then:

$$\begin{aligned} w &= \sum_{i=1}^{n}\alpha_i y_i \varphi(x_i), \\ f(x) &= w\cdot\varphi(x)+b = \sum\nolimits_{i=1}^{n}\alpha_i y_i \varphi(x_i)\cdot\varphi(x)+b \\ &= \sum\nolimits_{i=1}^{n}\alpha_i y_i K(x_i,x)+b. \end{aligned} \tag{6}$$

A widely used Gaussian kernel function is the radial basis function:

$$K(x,y) = \exp\left(-\gamma * \lVert \text{x-y}\rVert^2\right). \tag{7}$$

$\lVert \text{x-y}\rVert^2$ is the square of the Euclidean distance between observation x and observation y. γ is the bandwidth of the kernel function. When the radial basis function is utilized as a kernel function, the adjusted parameters are the bandwidth of the kernel function and the penalty term C. The optimal parameters are determined by grid searches.

G. CROSS VALIDATION AND TESTS

Five-fold cross validation is applied in this research. The major steps are as follows:

1. Split the sections in the training dataset into 5 partitions. Each partition is a fold;
2. Iterate training and testing 5 times. In each iteration, a different fold is chosen as test data for this iteration. The other four folds are combined to form training data for this iteration;
3. Evaluate performances resulted from the iterations.

Next, use a new testing dataset to implement another test. In the test, the data are newer data from which all the above data used in the cross validation are excluded. The optimal model can be justified through all the training, cross validation and tests. Data in the testing dataset are updated before writing this paper.

H. FRAMEWORK FOR PRESENTED TECHNOLOGIES

Models are built by inputting the prepared data, and results are acquired from the different models. Assuming the classification threshold is P_0, a section will be predicted to be in worse condition if the probability that defects occur in the section is P_0 or larger. Additionally, a section will be predicted to be in better condition if the probability that defects occur in the section is smaller than P_0. Under different thresholds, recall rates and the number of worse sections can be calculated. The workload in this research is defined as the number of sections that are predicted to be in worse condition and will need the arranged labor force and equipment as usual (i.e., heavy maintenance). A threshold that maximizes the recall rate (100% in the final tests is the goal) and minimizes the workload is the best choice (the computational speed is also tested). The following shows a summary of the framework for technologies presented.

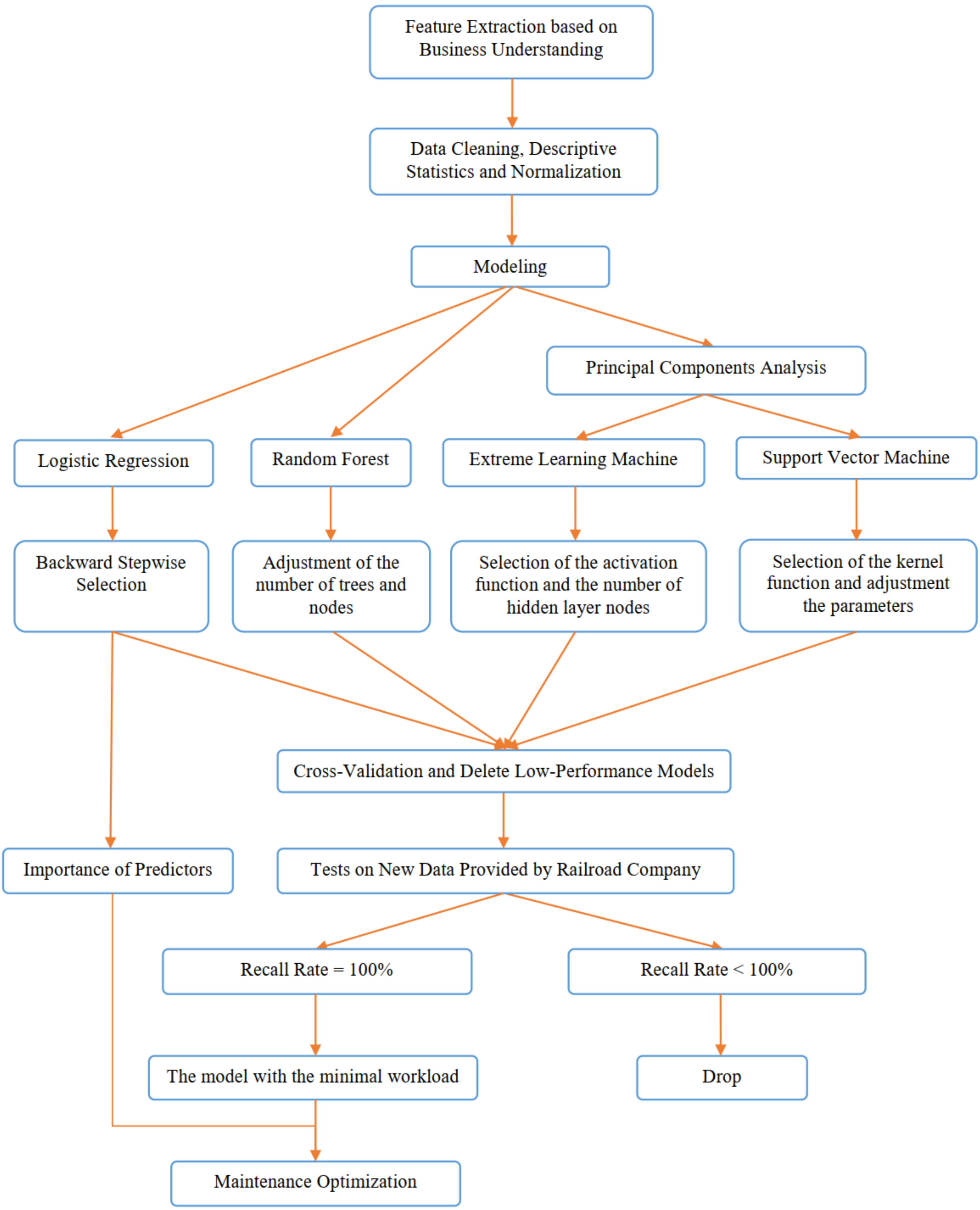

FIGURE 3. Framework for presented technologies.

III. RESULTS AND EVALUATION

A. EVALUATION METRICS AND PROCESS

Because it is impossible to find a model that works perfectly for both the recall rate and the workload, an optimal balance between them is a critical point. For safety reasons, it is necessary to find the rail sections in worse condition. Most railway managers require a 100% recall rate for the rail defects. However, as introduced previously, minimizing maintenance work is also one of our priorities. Therefore, the models are adjusted so that the results can embody an extremely high recall rate and an optimized workload. In addition, the models may be implemented efficiently.

To estimate the parameters of logistic regression, *P*-values are the key to determining whether the estimates can pass hypothesis tests. Then, the constructed model is optimized through backward stepwise selection, and hypothesis tests are also applied to the optimized model. The model with the smallest AIC value is chosen as the best logistic regression model [45]. The value of the workload expected to be minimized is decided by prediction precision at a given recall rate. The final model is further adjusted by finding the optimal threshold to balance the recall rate and the precision. For the random forest model, an optimal threshold is also found to satisfy the required balance between the recall rate and the precision. Additionally, the number of nodes in the hidden layer in ELM is adjusted based on the desired recall rate and precision to acquire the optimal model. For the SVM model with the radical basis function as its kernel function, the bandwidth of the kernel function and the penalty term are adjusted based on the desired recall rate and precision to acquire the optimal SVM model.

The parameters of the models are iteratively tuned to reach the best performance. The highest recall rate is a top priority in these adjustments. In the transportation industry, passenger safety is so important that all work should provide sufficient considerations to ensure safety. False negative predictions may lead to unexpected rail defects, which are threats to passengers and goods. As a result, the recall rate should be as high as possible. In addition, railroad companies would like to predict rail defects quickly. Thus, calculation efficiency is also taken into consideration.

Cross-validation is applied. All the data are real data from railroad companies. The training data and the validation data are generated and validated no later than September 2018. The testing data are generated and validated after September 2018. Again, the testing data are newer data not used in the modeling.

The final classification thresholds are determined by the recall rate and the workload from the testing dataset. Then, the optimal model is selected and confirmed.

B. DESCRIPTIVE ANALYSIS

1) DESCRIPTIVE STATISTICS

According to descriptive statistics, the centralization, discreteness and distribution are determined as follows:

TABLE 3. Descriptive statistics.

Variables	Mean	S.D.	Median	Trimmed	Min	Max	Skew	Kurtosis
Direction	1.51	0.54	1.00	1.49	1.00	3.00	0.32	-1.15
Line_type	1.29	0.45	1.00	1.24	1.00	2.00	0.92	-1.16
Speed_Grade	242.66	68.32	250.00	241.24	80.00	350.00	0.09	-0.70
Start_mile	322.43	200.40	319.17	320.99	2.00	662.60	0.03	-1.32
End_mile	321.60	201.73	316.66	319.97	0.32	664.59	0.04	-1.32
Pave_date	104.32	33.09	115.00	105.26	13.00	201.00	-0.30	0.93
L_temp	30.60	3.07	31.60	30.84	9.70	42.50	-1.63	8.17
R_temp	30.61	3.09	31.70	30.85	9.70	42.50	-1.61	7.90
Load	148.94	69.77	119.09	141.40	2.65	476.91	1.80	5.99
TQI	4.79	2.27	4.04	4.50	0.00	14.64	1.25	1.34
Curve_number	0.21	0.50	0.00	0.10	0.00	4.00	2.79	9.50
Curve_radius	0.05	0.19	0.00	0.01	0.00	2.81	7.00	72.09
Slope_number	0.54	0.97	0.00	0.32	0.00	6.00	2.11	4.85
Slope_value	0.04	1.47	0.00	0.00	-5.97	6.00	0.66	7.99
Grind_number	0.07	0.40	0.00	0.00	0.00	4.00	6.79	51.82
YDH_number	0.13	0.77	0.00	0.00	0.00	8.00	6.39	43.98
YDH_rain_number	0.01	0.20	0.00	0.00	0.00	6.00	27.54	792.19
YDH_temp	0.69	4.14	0.00	0.00	0.00	32.30	6.27	39.25
YDH_DDL	1.18	6.52	0.00	0.00	0.00	41.50	5.42	27.73
LRH_number	0.70	2.21	0.00	0.10	0.00	21.00	4.35	23.38
LRH_rain_number	0.07	0.48	0.00	0.00	0.00	6.00	7.98	74.05
LRH_left_model	0.15	0.36	0.00	0.05	0.00	2.00	2.26	3.92
LRH_right_model	0.15	0.39	0.00	0.05	0.00	2.00	2.45	5.41
LRH_weather	1.03	0.18	1.00	1.00	1.00	2.00	5.33	26.39
LRH_before_temp	2.43	6.84	0.00	0.37	-3.40	40.30	2.91	7.72
LRH_after_temp	2.33	6.66	0.00	0.33	-3.70	40.30	2.98	8.20
LRH_abs_temp	0.10	1.17	0.00	0.00	0.00	27.00	18.90	377.60
LRH_up_width	6.17	11.88	0.00	4.04	0.00	30.00	1.40	-0.02
LRH_low_width	6.12	11.79	0.00	4.01	0.00	30.00	1.40	-0.02
LRH_QGL	0.41	0.80	0.00	0.27	0.00	2.20	1.41	0.00
LRH_reaction_time	2.19	4.25	0.00	1.34	0.00	13.50	1.45	0.19
LRH_user_fraction	9.66	0.74	10.00	9.86	6.00	10.00	-2.07	3.12
LRH_quiet_time	2.67	5.19	0.00	1.61	0.00	17.00	1.48	0.29
Defect	0.08	0.28	0.00	0.00	0.00	1.00	3.01	7.09

2) COLLINEARITY ANALYSIS

A correlation analysis is conducted on all the potentially correlated variables. It is found that there are correlations among the predictors. The number of conditions of the correlation matrix of the predictors is 520889.8, which is larger than 1000, suggesting the existence of severe collinearity [46], [47]. As mentioned in the previous parts, for logistic regression, backward stepwise selection is used to solve the problem. For the SVM and ELM models, principal component analysis is utilized to eliminate the collinearity.

C. MODELING RESULTS

1) LOGISTIC REGRESSION

First, logistic regression is carried out with regard to all the predictors. The results are as follows:

TABLE 4. Logistic regression results with regard to all the predictors.

Variables	Estimate	Pr(>|z|)	Variables	Estimate	Pr(>|z|)
(Intercept)	8.0770	0.16927	YDH_temp	-1.3500	0.44768
Direction	-0.3892	0.13711	YDH_DDL	1.2590	0.18651
Line_type	-4.1850	0.00128	LRH_number	-0.0069	0.93437
Speed_Grade	-0.0218	0.00316	LRH_rain_number	0.3892	0.28759
Start_mile	-3.1060	0.51885	LRH_date	0.0000	0.89475
End_mile	2.0660	0.66921	LRH_left_model	-7.8160	0.85589
Pave_date	0.2965	0.34223	LRH_right_model	2.5440	0.02159 *
L_temp	-0.1684	0.84992	LRH_weather	0.2135	0.85615
R_temp	0.3621	0.68504	LRH_before_temp	203.9000	0.9888
Load	-0.2399	0.46483	LRH_after_temp	-199.1000	0.98876
TQI	-0.0942	0.74151	LRH_abs_temp	-35.7500	0.98852
Curve_number	0.5171	0.10984	LRH_up_width	9.0300	0.40351
Curve_radius	0.0515	0.7397	LRH_low_width	-3.0880	0.76643
Slope_number	-0.3259	0.11026	LRH-QGL	-2.4630	0.34508
Slope_value	0.2469	0.03139	LRH_reaction_time	-1.9140	0.04841 *
Grind_number	0.0583	0.85842	LRH_user_fraction	-0.0518	0.82934
YDH_number	-1.0460	0.66241	LRH_quiet_time	-0.8950	0.25848
YDH_rain_number	-4.1660	0.99325			

Table 4 presents the parameter estimations for the predictors. Certain P values are less than 0.05. This means that the corresponding coefficients can pass hypothesis tests at a confidence level of 95%. However, the coefficients of the other predictors cannot pass the tests because the corresponding P values are larger than 0.05.

Next, backward stepwise selection is added. After this selection, the model with the minimal AIC value is chosen.

After the backward stepwise selection, only 16 predictors are still in the model (see Table 5). Five of the P values of the predictors are larger than 0.1, and the remaining 11 are less than 0.1. This means that the coefficients of the 11 predictors (see Table 6) can pass hypothesis tests at a confidence level of 90%. Thus, they can be considered risk factors for rail defects. Experienced railroad engineers also confirmed that these 11 predictors may be critical for defect occurrence.

TABLE 5. Logistic regression results after backward stepwise selection.

Variable	Estimate	Pr(>|z|)	Variable	Estimate	Pr(>|z|)
(Intercept)	9.8732	0.000398	YDH_temp	-2.2099	0.320594
Direction	-0.3712	0.137879	YDH_DDL	1.0168	0.208723
Line_type	-4.9231	5.71e-06	LRH_rain_number	0.4102	0.034797
Speed_Grade	-0.0247	6.65e-05	LRH_left_model	-2.0912	0.065184
Start_mile	-1.0451	0.000538	LRH_right_model	2.2951	0.025185
R_temp	0.2113	0.147641	LRH_after_temp	-0.5387	0.020868
Curve_number	0.5809	0.022802	LRH_up_width	2.1683	0.012407
Slope-number	-0.2861	0.116865	LRH_reaction-time	-1.5505	0.070582 .
Slope_value	0.2376	0.034194			

TABLE 6. Importance of the predictors.

	Variable	Estimate	Absolute Value of the Estimate	Pr(>|z|)
1	Line_type	-4.92306681	4.92306681	5.71e-06
2	LRH_right_model	2.29507107	2.29507107	0.025185
3	LRH_up_width	2.16829058	2.16829058	0.012407
4	LRH_left_model	-2.09118894	2.09118894	0.065184
5	LRH_reaction_time	-1.55053519	1.55053519	0.070582
6	Start_mile	-1.04513724	1.04513724	0.000538
7	Curve_number	0.58088803	0.58088803	0.022802
8	LRH_after_temp	-0.53870521	0.53870521	0.020868
9	LRH_rain_number	0.41019538	0.41019538	0.034797
10	Slope-value	0.23764531	0.23764531	0.034194
11	Speed_Grade	-0.02467043	0.02467043	6.65e-05

Let $p =$ the probability of defect occurrence in a section; the term Odds is defined as [48]:

$$Odds = \frac{p}{1-p} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p} \quad (8)$$

For each predictor, a change of 0.1 units leads to a change in Odds of $\left(2.72^{0.1*\beta} - 1\right)$. We call this the odds rate.

TABLE 7. Odds rates.

Variable	Odds Rate	Variable	Odds Rate
Line_type	-0.388785111	Curve_number	0.059809106
LRH_right_model	0.257979807	LRH_after_temp	-0.052445213
LRH_up_width	0.242131752	LRH_rain_number	0.041872462
LRH_left_model	-0.188701229	Slope_value	0.024049157
LRH_reaction_time	-0.143630656	Speed_Grade	-0.002464003
Start_mile	-0.099237565		

According to Figure 4, here is an approximately linear relationship between the odds rate and the coefficients of the

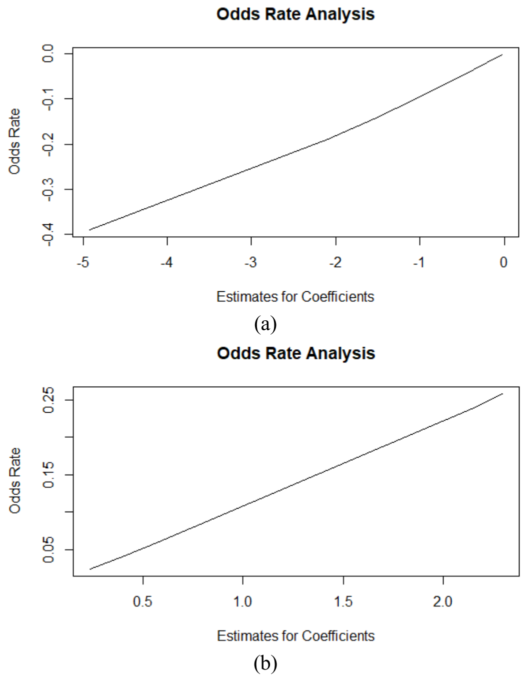

FIGURE 4. Odds rate analysis.

variables. The odds rate increases as the absolute values of the coefficients increase. Therefore, the larger the absolute values of the coefficients are, the greater the influence on the odds of the occurrence of defects.

In addition, AIC values were calculated. Table 8 shows that after the backward stepwise selection, the AIC value decreased to 510.94 from 539.19. This is a positive change that suggests that the model improved.

TABLE 8. Comparison between different logistic regression models.

Model	AIC
Regular Logistic Regression	539.19
Logistic Regression (BSS)	510.94

Note: BSS: Backward stepwise selection.

2) RANDOM FOREST

After iterative adjustments based on the training dataset, the number of trees is set to 320, and the number of nodes is chosen as 5.

3) EXTREME LEARNING MACHINE

First, principal component analysis is conducted on all the continuous predictors. The results are shown in Table 9. C.1, . . . , C.28 are 28 principal components.

According to Table 9, the cumulative proportion of the first 10 principal components is 87.8%. The next 18 principle components contribute little to the variance. Therefore, the first 10 principal components are selected for further analysis.

TABLE 9. Initial results of principal component analysis.

	S.D.	P.V.	C.P.		S.D.	P.V.	C.P.
C.1	3.0871	0.3404	0.3404	C.15	0.5806	0.0120	0.9711
C.2	1.8968	0.1285	0.4689	C.16	0.5193	0.0096	0.9807
C.3	1.6565	0.0980	0.5669	C.17	0.4263	0.0065	0.9872
C.4	1.5668	0.0877	0.6545	C.18	0.3273	0.0038	0.9911
C.5	1.1063	0.0437	0.6982	C.19	0.3225	0.0037	0.9948
C.6	1.0605	0.0402	0.7384	C.20	0.2875	0.0030	0.9977
C.7	1.0385	0.0385	0.7769	C.21	0.1451	0.0008	0.9985
C.8	1.0006	0.0358	0.8127	C.22	0.1334	0.0006	0.9991
C.9	0.9712	0.0337	0.8464	C.23	0.0974	0.0003	0.9994
C.10	0.9408	0.0316	0.8780	C.24	0.0894	0.0003	0.9997
C.11	0.9006	0.0290	0.9069	C.25	0.0801	0.0002294	0.9999625
C.12	0.7718	0.0213	0.9282	C.26	0.0312	0.0000347	0.9999972
C.13	0.6752	0.0163	0.9445	C.27	0.0078	0.0000022	0.9999993
C.14	0.6387	0.0146	0.9591	C.28	0.0043	0.0000007	1.0000000

Notes: S.D.: Standard deviation; P.V.: Proportion of variance; C.P.: Cumulative proportion.

The 10 principal components (in Table 10 and Table 11) replaced the corresponding variables with collinearity effects. The number of nodes in the ELM model is adjusted between 1 and 500. This adjustment is applied to ELM models with all the different activation functions presented in Table 2.

Then, the ELM models are tuned with the activation functions. The results show that 4 activation functions perform significantly better than the others. They are the symmetric saturation linear (satlins), rectifier linear unit (relu), triangular basin (tribas) and linear (purelin) functions. Table 12 shows the related parameters.

4) SUPPORT VECTOR MACHINE

To address the collinearity problem, the above PCA selection is also applied to the SVM model. After iterative adjustments based on the training dataset, the penalty item is set to 1, and the bandwidth of the kernel function is chosen as 8.

D. CROSS-VALIDATION, TEST AND MODEL SELECTION

Five-fold cross validation is applied. The average workload presented below is the average validated workload, and the average recall rate is the average validated recall rate.

All the models we created show very high recall rates and favorable workloads. Therefore, these models may perform well on the testing dataset which leads to the final model selection, and they have passed the cross-validation.

Then, the testing dataset is used to test the above models. The security rate term is defined as the number of predicted defect-free sections divided by the total number of testing sections.

For regular-speed rail, the recall rates are all 100% in the test, and ELM (satlins) shows the optimal workload. The performance of the ELM (satlins) is visualized.

For high-speed rail, ELM (relu) is dropped because its recall rate is under 100%. Then, the random forest model

TABLE 10. Parameters of selected principal components.

	C.1	C.2	C.3	C.4	C.5	C.6
Start_mile	0.157	0.16	0.384	0.142	0.319	0.158
End_mile	0.154	0.159	0.381	0.141	0.32	0.159
Pave_date	/	0.323	/	0.301	-0.15	0.336
L_temp	/	/	0.169	0.559	-0.155	-0.123
R_temp	/	/	0.168	0.559	-0.149	-0.133
Load	/	0.22	-0.359	/	-0.333	0.286
TQI	-0.215	-0.181	-0.126	0.175	-0.181	-0.155
Curve_number	-0.128	/	-0.234	0.241	0.279	/
Curve_radius	-0.117	/	-0.284	0.174	0.387	-0.105
Slope_number	-0.175	/	-0.25	0.217	/	0.158
Slope_value	/	/	/	/	/	0.439
Grind_number	/	/	-0.28	0.155	0.399	/
YDH_number	/	-0.48	/	/	/	0.226
YDH_rain_number	/	-0.158	/	/	/	0.508
YDH_temp	/	-0.467	/	/	/	/
YDH_DDL	/	-0.49	/	/	/	0.115
LRH_number	-0.235	/	/	-0.103	0.162	/
LRH_rain_number	-0.125	/	/	-0.123	0.336	-0.108
LRH_date	-0.287	/	/	/	/	/
LRH_before_temp	-0.263	/	/	/	/	0.108
LRH_after_temp	-0.259	/	/	/	/	/
LRH_abs_temp	/	/	/	/	0.189	0.321
LRH_up_width	-0.3	/	0.163	/	/	/
LRH_low_width	-0.3	/	0.163	/	/	/
LRH_QGL	-0.298	/	0.166	/	/	/
LRH_reaction_time	-0.297	/	0.162	/	/	/
LRH_user_fraction	0.284	/	-0.111	/	/	/
LRH_quiet_time	-0.296	/	0.165	/	/	/

shows the optimal workload. The performance of the random forest model is visualized.

The calculation speeds are fast enough for all the models. The models' performance on the testing dataset for regular-speed railways is shown in Table 14, Figure 6 and Figure 7. We can observe that ELM (satlins) significantly decreases the workload with a 100% recall rate. For this recall rate, ELM (satlins) shows the highest security rate and the smallest workload. Therefore, the ELM (satlins) model is the best selection for regular-speed rail under the business goals.

The models' performance on the testing dataset for high-speed railways is presented in Table 15, Figure 8 and Figure 9. We can observe that the random forest model significantly decreases the workload with a 100% recall rate. For this recall rate, the random forest model shows the highest security rate and the smallest workload. Therefore, the random forest model is the best selection for high-speed rail under the business goals.

TABLE 11. Parameters of selected principal components (Continued).

	C.7	C.8	C.9	C.10
Start_mile	/	/	0.103	0.184
End_mile	/	/	0.105	0.187
Pave_date	/	/	0.126	0.373
L_temp	-0.179	/	/	/
R_temp	-0.181	/	/	/
Load	0.153	/	0.104	0.277
TQI	/	/	/	/
Curve_number	0.167	/	/	/
Curve_radius	0.147	/	0.148	/
Slope_number	/	/	/	0.118
Slope_value	/	-0.694	-0.529	-0.136
Grind_number	0.15	-0.193	0.103	-0.178
YDH_number	/	/	/	0.124
YDH_rain_number	-0.291	/	0.578	-0.472
YDH_temp	0.14	/	-0.1	0.311
YDH_DDL	/	/	/	0.245
LRH_number	-0.257	/	/	0.109
LRH_rain_number	-0.413	-0.142	/	0.159
LRH_date	-0.216	/	/	/
LRH_before_temp	-0.289	0.171	-0.136	/
LRH_after_temp	-0.322	/	/	0.162
LRH_abs_temp	0.147	0.628	-0.457	-0.369
LRH_up_width	0.191	/	/	/
LRH_low_width	0.194	/	/	/
LRH_QGL	0.2	/	/	/
LRH_reaction_time	0.205	/	/	/
LRH_user_fraction	/	/	/	/
LRH_quiet_time	0.209	/	/	/

TABLE 12. ELM model parameters.

Model	ELM (satlins)	ELM (relu)	ELM(tribas)	ELM (purelin)
Nodes	8	16	213	5

TABLE 13. Cross-validation results.

Model	Average Workload	Average Recall Rate
LR with BSS	158.2	98.2%
RF	111	100%
ELM (satlins)	159	100%
ELM (relu)	163	100%
ELM (tribas)	160	100%
ELM (purelin)	176	100%
SVM	170	100%

Notes: LR: Logistic Regression, BSS: Backward Stepwise Selection, RF: Random Forest.

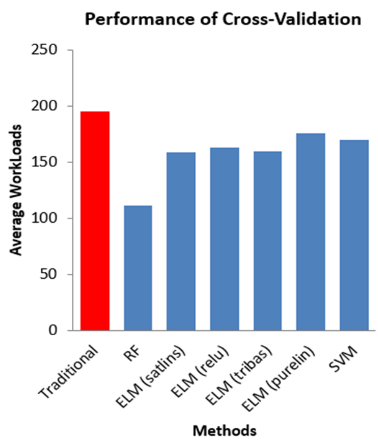

FIGURE 5. Performance evaluation based on cross-validation.

Moreover, the threshold of the optimal model for high-speed rail is significantly stricter than the threshold of

TABLE 14. Prediction performance on regular-speed rail.

Models	Workload	Recall	Threshold	Security Rate	Computational Time
ELM (satlins)	182	100%	8.43%	31.58%	<0.01s
ELM (tribas)	191	100%	3.692%	28.2%	<0.01s
ELM (relu)	209	100%	1.82%	21.43%	<0.01s
RF	228	100%	1.875%	14.29%	<0.01s
LR with BSS	239	100%	1%	10.15%	0.01s
ELM(purelin)	246	100%	3.81%	7.52%	<0.01s
SVM	257	100%	7.969%	3.38%	<0.01s

Notes: LR: Logistic Regression, BSS: Backward Stepwise Selection, RF: Random Forest.

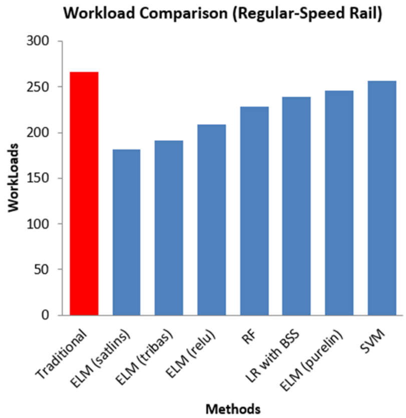

FIGURE 6. Performance evaluation based on the testing dataset (Regular-speed Rail).

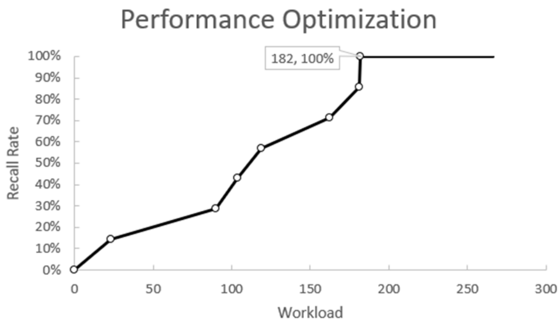

FIGURE 7. ELM (satlins) workload optimization.

TABLE 15. Prediction performance on high-speed rails.

Models	Workload	Recall	Threshold	Security Rate	Computational Time
RF	346	100%	4.063%	29.82%	0.01s
LR with BSS	372	100%	5.81%	24.54%	0.01s
ELM (tribas)	418	100%	3.446%	15.21%	<0.01s
ELM (satlins)	463	100%	5.64%	6.09%	<0.01s
SVM	476	100%	7.453%	3.45%	0.01s
ELM (purelin)	487	100%	4.604%	1.22%	<0.01s
ELM (relu)	337	87.5%	4.32%	31.64%	<0.01s

Notes: LR: Logistic Regression, BSS: Backward Stepwise Selection, RF: Random Forest.

the optimal model for regular-speed rail. This is good because of the higher safety requirements for high-speed rail.

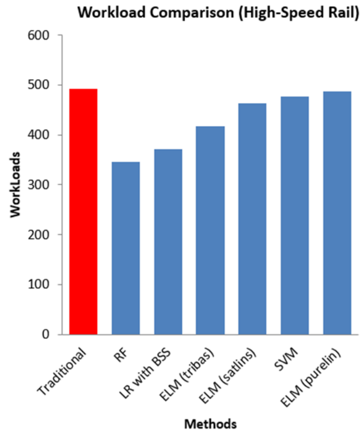

FIGURE 8. Performance evaluation based on the testing dataset (High-speed Rail).

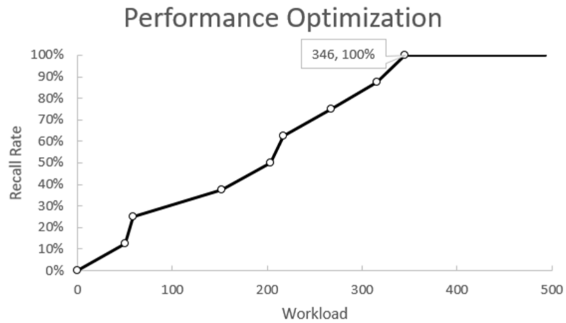

FIGURE 9. Random forest workload optimization.

Again, the above research also addresses the foundation of the railroad business: safety.

IV. DISCUSSION & CONCLUSIONS

The prediction methods and evaluation have been completed. The condition of a section can be predicted by the optimal models as follows: a better track condition in which no defects are predicted or a worse track condition in which defects are predicted to occur.

The findings in this paper provide important references for decision makers to design predictive maintenance. The models developed in this paper will assist engineers in terms of railway inspection, safety management, cost control, schedule optimization, etc.

Historically, the schedule for rail inspection was based on the assumption that all sections require frequent inspection to ensure safety. Through the new models proposed above, workloads will be rearranged, and lower workloads will be required for sections in better condition. Currently, welds from these better sections are inspected regularly by thousands of engineers, but the rail departments will be able to decrease the workloads by half based on our models. In China, it is estimated that more than 120,000 labor days per month (equal to hiring 4,000 workers to work 30 days) are needed to conduct related inspection work. According to the above tests, for regular-speed rail, the welds predicted to not have any defects for about a year (the period of the testing dataset) are 31.58% of the total welds, and this value is 29.82% for high-speed rail. The 100% recall rate means an extremely low probability for defect occurrence. According to our newest database, more than 95% of the defects are minor defects that do not require any repair. Therefore, the risk for the rail sections predicted to not have any defects is very low. Inspection work for the low-risk sections should be reduced significantly. These sections account for approximately 30% of the total sections. This suggests a massive cost reduction for weld inspection.

Additionally, 11 risk factors contributing to rail defects and their risk levels are identified. It is recommended that railroad companies pay more attention to these factors.

The inputs for the whole data mining process are data that are widely available in the daily operations of railroad companies. Additionally, the models have been integrated into one of our newest information systems for implementation. Data are updated, validated and prepared based on strict processes in accordance with business requirements.

Consequently, railway companies are expected to achieve enhanced management and operation with cost savings.

REFERENCES

[1] S. Alahakoon, Y. Q. Sun, M. Spiryagin, and C. Cole, "Rail flaw detection technologies for safer, reliable transportation: A review," *J. Dyn. Syst., Meas., Control*, vol. 140, no. 2, Feb. 2018, Art. no. 020801.

[2] Q. Hao, X. Zhang, Y. Wang, Y. Shen, and V. Makis, "A novel rail defect detection method based on undecimated lifting wavelet packet transform and Shannon entropy-improved adaptive line enhancer," *J. Sound Vibrat.*, vol. 425, pp. 208–220, Jul. 2018.

[3] X. Cao, W. Xie, S. M. Ahmed, and C. R. Li, "Defect detection method for rail surface based on line-structured light," *Measurement*, vol. 159, Jul. 2020, Art. no. 107771.

[4] Q. Wei, X. Zhang, Y. Wang, N. Feng, and Y. Shen, "Rail defect detection based on vibration acceleration signals," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I MTC)*, May 2013, pp. 1091–5281.

[5] V. R. Vijaykumar and S. Sangamithirai, "Rail defect detection using Gabor filters with texture analysis," in *Proc. 3rd Int. Conf. Signal Process., Commun. Netw. (ICSCN)*, Mar. 2015, pp. 1–6.

[6] Y. Min, B. Xiao, J. Dang, B. Yue, and T. Cheng, "Real time detection system for rail surface defects based on machine vision," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, p. 3, Dec. 2018.

[7] C. Ling, G. Jianqiang, G. Xiaorong, W. Zeyong, and L. Jinlong, "Research on rail defect detection system based on FPGA," in *Proc. IEEE Far East NDT New Technol. Appl. Forum (FENDT)*, Jun. 2016, pp. 211–216.

[8] F. Wu, Q. Li, S. Li, and T. Wu, "Train rail defect classification detection and its parameters learning method," *Measurement*, vol. 151, Feb. 2020, Art. no. 107246.

[9] Z. Xiong, Q. Li, Q. Mao, and Q. Zou, "A 3D laser profiling system for rail surface defect detection," *Sensors*, vol. 17, no. 8, p. 1791, Aug. 2017.

[10] C. Tastimur, H. Yetis, E. Akin, and M. Karaköse, "Rail defect detection and classification with real time image processing technique," *Int. J. Comput. Sci. Softw. Eng.*, vol. 5, no. 12, pp. 283–290, 2016.

[11] T. Heckel, H. M. Thomas, M. Kreutzbruck, and S. Rühe, "High speed non-destructive rail testing with advanced ultrasound and eddy-current testing techniques," in *Proc. 5th Int. Workshop NDT Experts, Brno Univ. Technol.*, 2009, pp. 101–109.

[12] Z. Su, A. Jamshidi, A. Núñez, S. Baldi, and B. De Schutter, "Integrated condition-based track maintenance planning and crew scheduling of railway networks," *Transp. Res. C, Emerg. Technol.*, vol. 105, pp. 359–384, Aug. 2019.

[13] R. Ahmad and S. Kamaruddin, "An overview of time-based and condition-based maintenance in industrial application," *Comput. Ind. Eng.*, vol. 63, no. 1, pp. 135–149, Aug. 2012.

[14] A. Jamshidi, S. Hajizadeh, Z. Su, M. Naeimi, A. Núñez, R. Dollevoet, B. De Schutter, and Z. Li, "A decision support approach for condition-based maintenance of rails based on big data analysis," *Transp. Res. C, Emerg. Technol.*, vol. 95, pp. 185–206, Oct. 2018.

[15] E. Fumeo, L. Oneto, and D. Anguita, "Condition based maintenance in railway transportation systems based on big data streaming analysis," *Procedia Comput. Sci.*, vol. 53, no. 1, pp. 437–446, 2015.

[16] B. Bergquist and P. Söderholm, "Data analysis for condition-based railway infrastructure maintenance," *Qual. Rel. Eng. Int.*, vol. 31, no. 5, pp. 773–781, Jul. 2015.

[17] S. Sharma, Y. Cui, Q. He, R. Mohammadi, and Z. Li, "Data-driven optimization of railway maintenance for track geometry," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 34–58, May 2018.

[18] Z. Su, A. Nunez, S. Baldi, and B. De Schutter, "Model predictive control for rail condition-based maintenance: A multilevel approach," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 354–359.

[19] H. Xu, J. Zhou, P. G. Asteris, D. Jahed Armaghani, and M. M. Tahir, "Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate," *Appl. Sci.*, vol. 9, no. 18, p. 3715, Sep. 2019.

[20] J. Zhong, Z. Liu, Z. Han, Y. Han, and W. Zhang, "A CNN-based defect inspection method for catenary split pins in high-speed railway," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 8, pp. 2849–2860, Aug. 2019.

[21] H. Huang, J. Xu, J. Zhang, Q. Wu, and C. Kirsch, "Railway infrastructure defects recognition using fine-grained deep convolutional neural networks," in *Proc. IEEE DICTA*, Dec. 2018, pp. 1–8,

[22] M. Guerrieri, G. Parla, and C. Celauro, "Digital image analysis technique for measuring railway track defects and ballast gradation," *Measurement*, vol. 113, pp. 137–147, Jan. 2018.

[23] H. Yao, C. Ulianov, and F. Liu, "Joint self-learning and fuzzy clustering algorithm for early warning detection of railway running gear defects," in *Proc. 24th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2018, pp. 1–8.

[24] A. Nunez, A. Jamshidi, and H. Wang, "Pareto-based maintenance decisions for regional railways with uncertain weld conditions using the Hilbert spectrum of axle box acceleration," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1496–1507, Mar. 2019.

[25] J. Z. Liu, D. Chen, and G. Zhao, "Track impact index method for evaluating track short wave irregularity of high-speed railway," *China Railway Sci.*, vol. 37, no. 4, pp. 34–41, 2016.

[26] A. Jamshidi, S. Faghih Roohi, A. Núñez, R. Babuska, B. De Schutter, R. Dollevoet, and Z. Li, "Probabilistic defect-based risk assessment approach for rail failures in railway infrastructure," *IFAC-PapersOnLine*, vol. 49, no. 3, pp. 73–77, 2016.

[27] K. G. Mercy and S. K. Srinivasa Rao, "A framework for rail surface defect prediction using machine learning algorithms," in *Proc. Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2018, pp. 972–977.

[28] P. Lesiak and P. Bojarczak, "Application of wavelets and fuzzy sets to the detection of head-checking defects in railway rails," in *Proc. Int. Conf. Transp. Syst. Telematics*. Berlin, Germany: Springer, 2010, pp. 327–334.

[29] C. S. Sun, J. Liu, and Y. Qin, "Intelligent detection method for rail flaw based on deep learning," *Zhongguo Tiedao Kexue/China Railway Sci.*, vol. 39, no. 5, pp. 51–57, 2018.

[30] W. Chen, W. Liu, K. Li, P. Wang, H. Zhu, Y. Zhang, and C. Hang, "Rail crack recognition based on adaptive weighting multi-classifier fusion decision," *Measurement*, vol. 123, pp. 102–114, Jul. 2018.

[31] N. Yao, X. Zhang, and K. Tao, "Research on correlation between track geometry and rail defects based on machine learning," in *Proc. World Congr. Railway Res.*, Tokyo, Japan, Oct. 2019. [Online]. Available: https://wcrr2019.org/cfp.html

[32] *Maintenance Regulations for Regular Speed Rail*, China Railway Corp., Beijing, China, 2019.

[33] *Implementation Rules for Regular Speed Rail Repair*, Shanghai Railway Admin. Corp., Shanghai, China, 2019.

[34] *Methodologies for High-Speed Rail Maintenance*, Shanghai Railway Admin. Corp., Shanghai, China, 2017.

[35] P. Horata, S. Chiewchanwattana, and K. Sunat, "Robust extreme learning machine," *Neurocomputing*, vol. 102, pp. 31–44, Feb. 2013.

[36] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[37] R. Bailey-Wood, C. M. Dallimore, S. A. Smith, and J. A. Whittaker, "Use of logistic regression analysis to improve prediction of prognosis in acute myeloid leukaemia," *Leukemia Res.*, vol. 8, no. 4, pp. 667–679, 1984.

[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[39] T. Kam Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.
[40] L. Breiman, "Statistical modeling: The two cultures," *Statist. Sci.*, vol. 16, no. 3, pp. 199–231, 2001.
[41] S. S. Dong and Z. X. Huang, "A brief theoretical overview of random forests," *J. Integr. Technol.*, vol. 2, no. 1, pp. 1–7, 2013.
[42] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 2, Jul. 2004, pp. 985–990.
[43] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.
[44] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[45] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. Sel. Papers HirotuguAkaike*. New York, NY, USA: Springer, 1988, pp. 199–213.
[46] J. M. Chambers and T. J. Hastie, "Linear models," in *Statistical Models*. Belmont, CA, USA: Wadsworth, 1992, ch. 4.
[47] E. Anderson, Z. Bai, C. Bischof, *LAPACK Users' Guide*, 3th ed. Philadelphia, PA, USA: SIAM, 1999. [Online]. Available: http://www.netlib.org/lapack/lug/lapack_lug.html
[48] (2013). *StatsProf*. [Online]. Available: http://logisticregressionanalysis.com/885-logistic-regression-odds-ratio-the-math/

NAN YAO was born in Beijing, China, in 1985. He received the B.S. degree from Fudan University, China, and the M.B.A. and M.S. degrees in information systems from Johns Hopkins University. He is currently a Data Scientist with China Academy of Railway Sciences Corporation Ltd. His main duty is to analyze data and to extract value from it, creating values through mining data from large and complex databases related to 139 000 km of railroads. He has ten years work experience in data mining. As a main Speaker, he has presented data mining results at the highest level world congress in railroad research at Tokyo, Japan, in 2019. He won the First Prize, which is awarded by the China Railway Society, for his data science work.

YUEJUN JIA was born in Zhejiang, China, in 1970. He received the master's degree in transportation engineering from Tongji University, China. He has worked as an Engineer at China Railway Shanghai Bureau Group Company Ltd. He has been engaged in maintenance technology and professional management for a long time. He has extensive experience in the production operation, rail flaw detection, organizing maintenance, inspection, and repair. He plays a key role in the reform for industrial supply integration. Moreover, he presided over the completion of the Severe-Broken Rail Monitoring System and GPC10-Type 1 Rail-Straightness Measuring Instrument and the Testing Platform, which awarded the Second Prize from the Shanghai Bureau to address Science and Technology Progress, in 2016. Because of the above accomplishments, he was elected as a Professional Top-Notch Talent of Railway Corporation and the Discipline Leader of Group Company, in 2019.

KAI TAO was born in Ningxia, China, in 1984. He received the master's degree in railroad engineering from Beijing Jiaotong University, China. He is currently the Director of the Data Center for Railroad Infrastructure of China Railway and an Associate Researcher with the Infrastructure Inspection Institute, China Academy of Railway Sciences Corporation Ltd. In the last ten years, he finished many national projects focusing on railway informatization.