Predicting the Severity of Accidents at Highway Railway Level Crossings of the Eastern Zone of Indian Railways using Logistic Regression and Artificial Neural Network Models

Anil Kumar Chhotu
Civil Engineering Department, National Institute of Technology Patna, India | Civil Engineering Department, Motihari College of Engineering, Motihari, India
anilc.phd19.ce@nitp.ac.in (corresponding author)

Sanjeev Kumar Suman
Civil Engineering Department, National Institute of Technology Patna, India
sksuman@nitp.ac.in

Received: 5 February 2024 | Revised: 17 February 2024 | Accepted: 4 March 2024

 DOI: https://doi.org/10.48084/etasr.7011

ABSTRACT

Road-railroad level crossing accidents pose serious safety risks to road users, and their significant increase requires more research efforts to propose substitute solutions. Such a solution must consider the impact of intersection geometry, user perception, traffic characteristics, driver behavior, environment, and seasonal variations on accidents. This study explores the considerable number of such accidents and develops a predictive model using all the factors that influence them. For these objectives, data were collected from databases maintained by the zonal head office of the East Central Railway (ECR) in India. Data included 175 level crossings that experienced at least one accident between 2006 and 2021 in the ECR region. This study presents two accident prediction models using logistic regression and ANN for the predominant factors of accidents in the ECR zone of Indian railways. The accuracy of fatal accident prediction was 96% for logistic regression and 98% for ANN.

Keywords-railroad level crossing; ANN; logistic regression; prediction model

I. INTRODUCTION

The Indian railway network is the world's third-largest, with a total route length of 68,442 km. It has 7,172 stations and serves more than 13,500 passengers in 9,100 freight trains daily [1]. Based on statistical data provided by Indian railways, India had a recorded total of 685 level crossing accidents between 2006 and 2021, among which 611 occurred at unmanned level crossings and 74 at manned level crossings. These accidents were responsible for 2,639 deaths and 4,991 non-fatal injuries [2]. Traffic accidents contribute to 43% of total accidents in India [3-4]. Some of the major reasons for the high rate of accidents are: (1) India is the second-largest populated country in the world, amplifying its volume exposure to highway traffic, (2) cars are the main means of transportation for most population, (3) India has several urban regions, (4) Most of the freight and traffic take place in big cities. These data show that great efforts are needed to reduce the threat of accidents that cause deaths and serious injuries in Indian cities. Accidents at unmanned level crossings for the period 2006-2020 show decreasing trends due to the new safety policies of the government. This study aims to analyze the factors that contribute in accidents in the east-central zone of Indian railways. Accident data were collected through the East Central Railways zone office, Hazipur, from 2006 to 2021. These data include 175 level crossings with at least one accident between 2006 and 2021. This study examines a variety of factors that differ from previous studies, such as geometric, vehicle-specific, terrain characteristics, environmental factors, driver characteristics. This study uses Logistic Regression (LR) and an Artificial Neural Network (ANN) model, which are soft computing tools. In particular, LR is one of the most important analytical tools in the social and natural sciences.

II. LITERATURE REVIEW

The intersections of highways and roadways can be manned or unmanned level crossings, where manned level crossings are closed when trains cross the section. However, unmanned level crossings are always open for vehicle movement. At unmanned crossings, drivers must make decisions based on the position of the train. These crossing points pose serious dangers not only for vehicles or trains but also for people who live nearby.

Several studies have attempted to analyze the various reasons for accidents that occur at highway-rail level crossings. In [5], US data between 2009 and 2014 were used to investigate accidents at private Higway-Railroad Level Crossings (HRLCs), showing that more accidents occurred at HRLCs without warning devices. In [6], data from the Federal Railroad Administration were examined for approximately 26,000 accidents at level crossings in the USA between 2002 and 2011, showing that more collisions occurred during the day than at night. This study also showed that accidents at level crossings appear to be significantly higher during the morning peak, evening peak, and evening off-peak compared to other times of the day. In [7], psychological factors that influence driver interruption and lack of concentration within the Australian continent's rail corridor were investigated, showing that excessive stress, disaffection, and cognitive flexibility are some communicative factors that can increase distraction and risks in driving. In [8], the recurrence of mortalities, the period of the accident, and the characteristics of individuals who died in train accidents on Finland's railways between 2005 and 2009 were examined. This study showed that pedestrians are more frequently involved in accidents. The US Department of Transportation identified human factors collectively as the main causes of collisions at HRLCs and investigated applicable action plans to address safety issues at level crossings [9]. Great attention was paid to education, engineering, and enforcement, which could reduce deaths in such cases. In [10], the severity of crashes involving two vehicles at unsigned intersections was predicted using ANNs, and the results showed very high accuracy compared to other statistical methods. In [11], a nonlinear hybrid model outperformed all other models in terms of prediction accuracy. In [12], pedestrian accidents in HRLCs were examined by dividing their severity into three categories: minor injury, major injury, and death. In [13], an ANN was more accurate compared to SVM, LR, and random forest.

HRLC accidents occur mainly due to drivers' perceptions of approaching trains, as there is a risk of error due to poor judgment of the train's location [14]. In [15], HRLCs were examined to determine to what extent the system approach has been used to adequately assess safety issues. This study stated that none of the previous studies used the systems approach to understand driver behavior at HRLC and investigated whether it could help select price-style augmentations at rail crossings. Uncertainties and errors for major road users were evaluated in a secure HRLC. Numerous infractions by both drivers and pedestrians were observed in the examined HRLC. This study also stated that a lack of proper planning could increase the number of violations in fully operational HRLCs. In [16], additional countermeasures were proposed to reduce the risk level in HRLCs [16].

Most reviewed studies were conducted in developed countries with high levels of education, low population density, and high per capita income. In addition, such countries use advanced infrastructure and technologies in RRLCs to prevent accidents. Therefore, it is necessary to investigate the ground realities of developing countries, such as India, where the population density is high and RRLCs are not equipped with recent technology to prevent accidents. This study developed an accident prediction model using most of the alarming factors for accidents in Indian RRLCs.

III. METHODOLOGY

A. Study Area and Data Collection

This study investigated the East Central Railway (ECR) zone of India, which is one of the 18 railway zones in the country. ECR has an extensive array of 5,402.693 track km and 3,707.988 route km. Data were collected from the zonal head office database from 2006 to 2021. This study collected data on level-crossing accidents through the Right to Information (RtI) Act. Some additional data were collected from Google Images, site visits, social networks, newspapers, and telephone communication with people from the area concerned. Data were collected based on the influencing factors for an accident at the level crossing. These data contain the location of the accident, time, date, category of the train involved, type of vehicle, number of fatal accidents, type of injuries, type (manned or unmanned) of level crossing, surface of the road at the level crossing, annual average daily traffic, and number of railway tracks. The main independent variables are the number of tracks, day and night accidents, manned and unmanned level crossings, season (summer or winter), train speed, warning device, AADT, type of road surface, and rural or urban areas.

B. Model Selection

1) Logistic Regression

LR techniques are used to create computational models of probabilistic systems to estimate future events. These probabilistic models do not impose any restrictions on the distributions of explanatory variables or predictors [18]. The logistic response function is expressed by the likelihood p that a binary response variable Y equals to 1 when the input variable X equals to x.

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_n X_{n)}}} \quad (1)$$

This function is non-linear and depicts an S-shaped curve. The predictor coefficients are denoted by β.

2) Artificial Neural Networks (ANNs)

An ANN is a popular machine learning method for making predictions. Feedforward and feedback ANNs are the two main architecture varieties. Neural models have been developed by analyzing feedforward or multilayer networks. The scaled conjugate approach is used to optimize the model. The acquired data are divided into two sets: training and testing. The model was trained with 70% of the data and then tested with the remaining 30%.

C. Primary Data Analysis

This study focuses mainly on the characteristics of RRLCs that were involved in accidents between 2006 and 2021. In this period, a minimum of one accident occurred at 75 of the 655 uncontrolled HRLCs in the east-central division, to a total of 175. According to the analysis of the collected data, the frequency of HRLC accidents increased from 2006 to 2014 and then declined during the subsequent seven years. The maximum number of deaths was observed in 2011, while there

were no incidents in 2019 and 2020. The major efforts of the Indian government in improving road safety and strategic planning have led to a substantial decrease in accidents. By 2022, the Indian Railways intended to have removed all unmanned crossings from the country's main roadways. Environmental factors, traffic conditions, driver traits, type of alert system in use, surface of the intersection area, light availability, and type of HRLC play a role in the number of casualties in the east-central zone. Most accidents (86.7%) from 2006 to 2021 involved non-gated HRLCs. At the most dangerous intersections, 67 HRLCs do not have enough illumination. Most unattended crossings have improperly kept crossing bucks or stop signs, while some of these signs are cracked or have a blurred color scheme. About 20% of HRLCs do not have adequate protection, which is a contributing factor to accidents.

D. Preparation of Model Data and Analysis

This study defined fatal and non-fatal accidents as dependent variables to establish a predictive model of the railroad level crossing accident. A fatal accident is coded with $y = 1$ and a non-fatal accident with $y = 0$. Table I shows all coded independent variables. Most variables were coded on a binary scale.

TABLE I. DETAILS OF INDEPENDENT VARIABLES

Variable	Abbreviation	Coded value
No of tracks (Railways)	T_R	0= One track 1=Two tracks
Rural or urban Area	A_{UR}	0= Rural area 1=Urban area
Summer and winter season	S_{SW}	0=Summer season 1=Winter season
Day and night	D_{DN}	0=Night 1=Daytime
Manned and unmanned level crossing	LC_{MU}	0=Manned Level crossing 1=Unmanned level crossing
Weekdays and Weekend	W_{WWD}	0=Weekdays 1=Weekend
Speed of the train	V	Actual speed
Warning device	W_{YN}	0=Not installed properly 1=Installed properly
Surface type	S_{YN}	0=Not good 1=Good
Average daily traffic	A_{ADT}	Actual AADT

IV. RESULTS AND DISCUSSION

A. Logistic Regression Model

A binary LR model was created using IBM SPSS version 22, and was used for the analysis. Table II shows the results. Table III shows that the proposed model has an accuracy of 0.96, which is exceptionally close to 1. Similarly, the sensitivity and specificity are 0.98 and 0.07, which are very close to 1 and 0, respectively. The Area Under the Curve (AUC) for the MLP model is 0.93, which is more than 0.9 and is defined as outstanding as shown in Table V.

1) Logistic Regression Model Validation

The LR model was validated using pseudo-R^2 statistical tests, which also assessed the fitness of the proposed model and provided satisfactory results for all tests. Table IV shows all the validation test results.

TABLE II. LOGISTIC REGRESSION RESULTS - VARIABLES

Variable	Estimates	S.E.	*p*-value
No of tracks	3.502	1.697	.039
Rural/Urban	5.174	1.958	.008
Winter/Summer	4.080	1.721	.018
Day/Night	2.830	1.365	.038
Manned/Unmanned crossings	3.306	1.545	.032
Weekend/Weekdays	5.379	2.231	.016
Speed	.072	.032	.026
Warning device	1.234	1.338	.357
AADT	1.432	1.573	0.21
Intercept	-18.60	5.147	.000

TABLE III. LOGISTIC REGRESSION RESULTS - FATALITIES

Confusion matrices						
	Non-fatal	Fatal	Accuracy	Sensitivity	Specificity	AUC
Non-fatal	71	1	0.96	0.98	0.07	0.93
Fatal	6	97				
Permissible value			1.00	1.00	0	0.9-1.0

TABLE IV. PSEUDO R^2

Pseudo R^2	Proposed model value	Permissible value
-2 log-likelihood (-2LL) test	0.105	0-1 (value toward zero more perfect model)
Cox and Snell R square	0.963	1
Nagelkerke R Square	0.973	1
McFadden's pseudo-R-square	0.321	0.20-.40

TABLE V. AUC VALUE

Area under curve=0.5	No distinction,
AUC between 0.5 to 0.6	Poor distinction
AUC between 0.6 to 0.7	Satisfactory distinction.
AUC between 0.7 to 0.8	Magnificent distinction
AUC between 0.9 to 1.0	Outstanding distinction

B. Artificial Neural Network

An analysis of the Multi-Layer Perception (MLP) model was performed in the IBM SPSS 22 software environment. The activation function is the hyperbolic tangent for the input layer and SoftMax for the output layer. This analysis used a scaled conjugate gradient as an optimization algorithm. Only one intermediate layer was used for this model. 70% of the data was used for training and 30% for testing. Table VI shows that the overall training accuracy was 97.6% and the testing accuracy was 98.0%. The AUC for the MLP model was 0.985, which is more than 0.9, as validated by Table V, and thus the model can distinguish between fatal and non-fatal accidents very well.

TABLE VI. ANN RESULTS

Confusion matrices						
Accident	Training		Testing		Accuracy	
	Fatal	Non-Fatal	Fatal	Non-Fatal	Training	Testing
Fatal	71	0	30	1	100.0	96.8
Non-Fatal	3	51	0	19	94.4	100.0

1) Sensitivity Analysis

Sensitivity analysis was performed to determine which of the potential factors was the most influential. An ANN model was developed and the connection weights were interpreted using the calculations suggested in [24]. This theory was also used in [25] in civil engineering. With the help of this analysis, the Relative Importance (RI) of each independent variable was identified. The variables were ranked by decreasing order of their corresponding RI values, as shown in Table VII. As the ranks were assigned to each variable, it was observed that the variable 'speed of the train' (V) had the highest RI of 31.10%. The 'level crossing type' (LC_{MU}) variable was the second most important. Finally, the 'week and weekend day traffic' (W_{WWD}) had the lowest impact on model output.

TABLE VII. RELATIVE IMPORTANCE OF VARIABLES

Variable	V	LC_{MU}	D_{DN}	A_{UR}	S_{SW}
RI	31.1	14.5	12.9	6.6	6.1
Rank	1	2	3	4	5
Variable	W_{YN}	T_R	A_{ADT}	S_{YN}	W_{WWD}
RI	5.6	5.5	5.4	3.6	3.0
Rank	6	7	8	9	10

C. Probability of Accident Prediction Comparison

The predictions of the two models of the probability of an accident were compared. The ANN prediction model was more accurate than the LR model. Fatal accidents are predicted for a probability of more than 0.5. However, non-fatal accidents are predicted with a probability of less than 0.5, as shown in Figure 1. The reference line is taken as the probability line 0.5. This figure also shows the prediction errors.

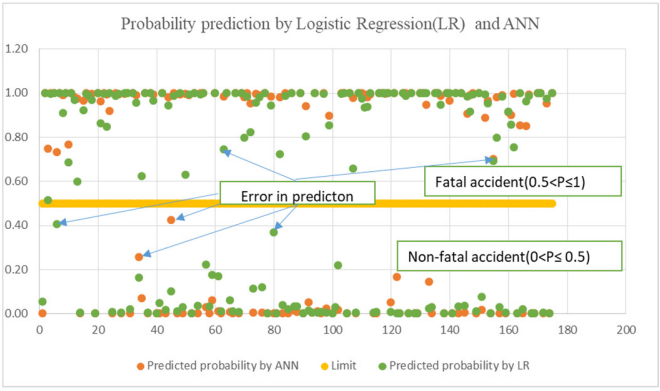

Fig. 1. Comparison of prediction between logistic regression and ANN.

V. CONCLUSION

This study demonstrates a predictive accuracy of more than 96% in determining the severity of level-crossing accidents using both methods. ANN had higher accuracy compared to LR. It was also shown that the extent of driver injuries in HRLCs is greatly affected by factors such as train speed, driving conditions throughout the day and night, weather conditions, location (rural or urban), number of railway tracks, and type of pavement surface at level crossings. According to the proposed approach, certain criteria, such as the presence of signs, road markings, and average daily traffic volume, are deemed irrelevant in the prediction of accidents. The sensitivity study reveals that train speed has the highest relative impact (31.1) on accidents, while weekday and weekend accidents have the lowest relative significance (3.0). Certain variables, namely driver and pedestrian behaviors, stopping sight distance, delay at the crossing, as well as peak hour factors, have the potential to be integrated into future research. Enhancing driver and pedestrian understanding of safety rules can reduce accidents at HRLCs in India. The government should enforce stringent sanctions on drivers who commit traffic violations at intersections, such as HRLCs. Furthermore, it was observed that most unmanned level crossings exhibit evidence of wear and tear or damaged road signs, necessitating proper maintenance. Some limitations of this study are listed below:

- Some important variables, such as sight distance, age of the driver, and weather conditions were not included in this study.
- All variables were coded in binary form.
- This study did not include property loss.
- The accuracy of the models is based only on the selected variables.
- The scope of this study is limited to similar traffic conditions as in the study area.

REFERENCES

[1] "Railway Networks of India." knowIndia.net, http://knowindia.net/rail.html.

[2] "Ministry of Railways (Railway Board)." https://indianrailways.gov.in/railwayboard/view_section.jsp?lang=0&id=0,1,304,366,554.

[3] National Crime Records Bureau - Ministry of Home Affairs, "Accidental Deaths and Suicides in India 2022." https://data.opencity.in/dataset/accidental-deaths-and-suicides-in-india-2022.

[4] K. Haleem and A. Gan, "Contributing factors of crash injury severity at public highway-railroad grade crossings in the U.S.," *Journal of Safety Research*, vol. 53, pp. 23–29, Jun. 2015, https://doi.org/10.1016/j.jsr.2015.03.005.

[5] K. Haleem, "Investigating risk factors of traffic casualties at private highway-railroad grade crossings in the United States," *Accident Analysis & Prevention*, vol. 95, pp. 274–283, Oct. 2016, https://doi.org/10.1016/j.aap.2016.07.024.

[6] W. Hao, C. Kamga, and D. Wan, "The effect of time of day on driver's injury severity at highway-rail grade crossings in the United States," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 3, no. 1, pp. 37–50, Feb. 2016, https://doi.org/10.1016/j.jtte.2015.10.006.

[7] A. Naweed, "Psychological factors for driver distraction and inattention in the Australian and New Zealand rail industry," *Accident Analysis & Prevention*, vol. 60, pp. 193–204, Nov. 2013, https://doi.org/10.1016/j.aap.2013.08.022.

[8] A. Silla and J. Luoma, "Main characteristics of train–pedestrian fatalities on Finnish railroads," *Accident Analysis & Prevention*, vol. 45, pp. 61–66, Mar. 2012, https://doi.org/10.1016/j.aap.2011.11.008.

[9] G. C. Cothen, "Role of Human Factors in Rail Accidents | US Department of Transportation." https://www.transportation.gov/testimony/role-human-factors-rail-accidents.

[10] S. A. Arhin and A. Gatiba, "Predicting Injury Severity of Angle Crashes Involving Two Vehicles at Unsignalized Intersections Using Artificial Neural Networks," *Engineering, Technology & Applied Science*

Research, vol. 9, no. 2, pp. 3871–3880, Apr. 2019, https://doi.org/10.48084/etasr.2551.

[11] Y. Kassem, H. Camur, M. T. Adamu, T. Chikowero, and T. Apreala, "Prediction of Solar Irradiation in Africa using Linear-Nonlinear Hybrid Models," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11472–11483, Aug. 2023, https://doi.org/10.48084/etasr.6131.

[12] A. Khattak and L.-W. Tung, "Severity of Pedestrian Crashes at Highway-Rail Grade Crossings," *Journal of the Transportation Research Forum*, vol. 54, no. 2, Jun. 2015, https://doi.org/10.5399/osu/jtrf.54.2.4291.

[13] F. Mlawa, E. Mkoba, and N. Mduma, "A Machine Learning Model for detecting Covid-19 Misinformation in Swahili Language," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10856–10860, Jun. 2023, https://doi.org/10.48084/etasr.5636.

[14] P. M. Salmon, G. J. M. Read, N. A. Stanton, and M. G. Lenné, "The crash at Kerang: Investigating systemic and psychological factors leading to unintentional non-compliance at rail level crossings," *Accident Analysis & Prevention*, vol. 50, pp. 1278–1288, Jan. 2013, https://doi.org/10.1016/j.aap.2012.09.029.

[15] G. J. M. Read, P. M. Salmon, and M. G. Lenné, "Sounding the warning bells: The need for a systems approach to understanding behaviour at rail level crossings," *Applied Ergonomics*, vol. 44, no. 5, pp. 764–774, Sep. 2013, https://doi.org/10.1016/j.apergo.2013.01.007.

[16] G. S. Larue, A. Naweed, and D. Rodwell, "The road user, the pedestrian, and me: Investigating the interactions, errors and escalating risks of users of fully protected level crossings," *Safety Science*, vol. 110, pp. 80–88, Dec. 2018, https://doi.org/10.1016/j.ssci.2018.02.007.

[17] A. Keramati, P. Lu, D. Tolliver, and X. Wang, "Geometric effect analysis of highway-rail grade crossing safety performance," *Accident Analysis & Prevention*, vol. 138, Apr. 2020, Art. no. 105470, https://doi.org/10.1016/j.aap.2020.105470.

[18] F. E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, NY, USA: Springer, 2001.

[19] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, https://doi.org/10.1016/j.patrec.2005.10.010.

[20] S. Eksteen and G. D. Breetzke, "Predicting the abundance of African horse sickness vectors in South Africa using GIS and artificial neural networks," *South African Journal of Science*, vol. 107, no. 7, pp. 1–8, Jan. 2011, https://doi.org/10.10520/EJC97173.

[21] S. Menard, *Applied Logistic Regression Analysis*. Thousand Oaks, CA, USA: SAGE, 2002.

[22] K. A. Pituch and J. P. Stevens, Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS, Sixth Edition. New York, NY, USA: Routledge, 2015.

[23] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, https://doi.org/10.1016/j.patrec.2005.10.010.

[24] D. W. H. Jr and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons, 2004.

[25] M. A. Shahin, H. R. Maier, and M. B. Jaksa, "Predicting Settlement of Shallow Foundations using Neural Networks," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 128, no. 9, pp. 785–793, Sep. 2002, https://doi.org/10.1061/(ASCE)1090-0241(2002)128:9(785).

AUTHORS PROFILE

Anil Kumar Chhotu is currently working as an assistant professor in the Department of Civil Engineering in the Motihari College of Engineering. He is also pursuing a Ph.D. from the National Institute of Technology, Patna. He has 9 years of teaching and research experience. He has more than 15 publications in different national and international reputed journals. He research area is road and rail safety, planning, and sustainable materials.

Sanjeev Kumar Suman is currently working as an Associate Professor in the Department of Civil Engineering at NIT Patna, Patna, India. He has eighteen years teaching and research experience. He holds a Ph.D. degree in transportation engineering. His research interests include road safety, sustainable and waste pavement materials, NDT pavement evaluation, pavement management systems, and construction quality control.