INTELLIGENT ROAD ACCIDENT ANALYSIS THROUGH MACHINE LEARNING ALGORITHMS

Deepali[1], Dr. Niresh Sharma[2]

Research Scholar, Department of Computer Science and Engineering, RKDF Institute of Science & Technology Bhopal (M.P.) [1]

Professor, Department of Computer Science and Engineering, RKDF Institute of Science & Technology Bhopal (M.P.) [2]

ABSTRACT

Road traffic accidents remain a persistent global public health crisis, disproportionately affecting developing nations including India, where over 461,000 accidents were recorded in 2022, claiming 168,491 lives. This empirical study investigates the application of machine learning algorithms to road accident data for severity classification, risk factor identification, and predictive modeling, using a compiled dataset of 52,814 accident records drawn from Indian national highway and urban road networks spanning 2018 to 2022. Seven machine learning algorithms were evaluated including logistic regression, decision trees, random forests, support vector machines, Naive Bayes, XGBoost, and long short-term memory neural networks, alongside an ensemble stacking model. Experimental results demonstrate that XGBoost achieved the highest individual accuracy of 91.3 percent with an AUC-ROC of 0.953, while the ensemble stacking model achieved 93.7 percent accuracy and an F1-score of 0.911, representing a statistically significant improvement over traditional statistical classifiers. Feature importance analysis identified vehicle speed, road surface condition, light condition, junction type, and alcohol impairment as the five most predictive variables for fatal accident outcomes. Severity analysis across road types revealed that rural roads carry the highest fatality rate at 11.3 percent despite lower overall accident volumes, underscoring the need for targeted rural safety interventions. Compared to analogous international studies, the ensemble model achieves competitive performance, placing it among the top three reported accuracy values in the published literature. These findings provide data-driven evidence for ML deployment within India's intelligent transportation infrastructure and offer a replicable empirical framework for road safety analysis in data-constrained environments.

KEYWORDS: *Road accident analysis[1], machine learning[2], accident severity classification[3], XGBoost[4], ensemble learning[5], feature importance[6], Indian road safety[7].*

1. INTRODUCTION

1.1 Road Accident Problem: Scale and Urgency

Road traffic accidents represent one of the most consequential and preventable public health challenges of the contemporary era, generating massive human, economic, and social costs across all world regions. According to the World Health Organization, approximately 1.35 million people die in road traffic crashes each year, with an additional 20 to 50 million sustaining non-fatal injuries of varying severity. The global economic cost of road accidents is estimated to exceed USD 1.8 trillion annually, representing 3 percent of most countries' gross domestic product. In the Indian context, the problem is particularly acute. Data published by the Ministry of Road Transport and Highways reveal that India accounts for approximately 11 percent of global road accident fatalities despite possessing only 1 percent of the world's vehicle fleet, an alarming disproportion reflecting the combined effect of infrastructure deficits, enforcement gaps, driver behavior issues, and emergency response limitations. The five-year period from 2018 to 2022 saw 2,155,928 total accidents recorded across the country, with cumulative fatalities exceeding 756,707 lives, as detailed in Table 1 of this paper. The compounding challenge is that accident causation in India is multi-factorial and highly context-dependent, varying substantially across national highways, state roads, urban arterials, and rural unclassified roads. This contextual complexity demands analytical approaches capable of capturing non-linear interactions among dozens of contributing variables simultaneously, a requirement that conventional statistical methods struggle to satisfy.

1.2 Machine Learning as an Analytical Paradigm

Machine learning offers a compelling analytical paradigm for road accident analysis by providing the capacity to identify complex, non-linear patterns in large, heterogeneous datasets without requiring a priori specification of functional forms or distributional assumptions. The field has progressed through several generations of increasingly powerful methods: from early decision tree and logistic regression approaches in the 1990s to ensemble methods such as random forests and gradient boosting in the 2010s, and most recently to deep learning architectures capable of processing image sequences, temporal traffic streams, and multi-source fusion data. Each generation has delivered measurable performance improvements on accident severity classification benchmarks, with the best contemporary models achieving accuracy rates above 93 percent on well-curated datasets. Critically, modern ML frameworks also provide feature importance and explainability tools, including SHAP value analysis and partial dependence plots, that allow practitioners to translate black-box model outputs into policy-actionable insights. The practical deployment of ML models in Indian road safety contexts has, however, lagged behind international literature, partly due to data standardization challenges and partly due to a shortage of empirically rigorous studies demonstrating model performance on Indian-specific accident datasets with transparent evaluation methodology.

1.3 Objectives and Contribution of This Study

This study addresses the identified gap by conducting a controlled empirical evaluation of seven machine learning algorithms on a compiled dataset of 52,814 Indian road accident records, with the following specific objectives: first, to benchmark the comparative predictive performance of classical ML algorithms and advanced ensemble methods on accident severity classification using standardized evaluation metrics; second, to quantify the relative importance of contributory features to accident severity outcomes through systematic feature importance analysis; third, to analyze severity distributions across road type categories to identify infrastructure-specific risk profiles; and fourth, to benchmark the present study's results against published international empirical studies to assess relative performance and highlight the transferability of findings to Indian conditions. This paper makes three primary contributions to the literature: it provides the first systematic multi-algorithm comparison on a large-scale compiled Indian accident dataset with full evaluation metric reporting; it presents a reproducible empirical framework combining preprocessing, class imbalance handling, feature selection, and model evaluation that can be applied to other regional accident datasets; and it delivers actionable findings on the dominant risk factors for fatal accidents in Indian road conditions that directly inform national safety policy priorities.

2. LITERATURE SURVEY

The empirical literature on machine learning applied to road accident analysis has grown substantially over the past two decades, progressing from isolated methodological experiments to systematic comparative studies on national-scale datasets. The foundational contributions of Karlaftis and Golias [1] using classification and regression trees on Greek accident data, and of Chang and Wang [7] demonstrating the superiority of back-propagation neural networks over logistic regression on Taiwanese crash records, established the empirical template that subsequent researchers would elaborate and extend. Delen, Sharda, and Bessonov [10] provided one of the first large-scale neural network analyses on US accident data, achieving 78.4 percent accuracy and identifying driver characteristics as primary severity determinants, while their simultaneous application of decision trees enabled a direct algorithm comparison that revealed the neural network's advantage on imbalanced data. The development of ensemble methods represented the most significant subsequent advance, with Iranitalab and Khattak [16] providing a controlled four-algorithm comparison on Kansas accident data that established random forest as the benchmark performer at 86.4 percent accuracy. The emergence of gradient boosting as the state of the art was confirmed by Sameen and Pradhan [17], whose XGBoost model achieved 91.2 percent accuracy on Malaysian highway data and introduced SHAP-based explainability analysis to the domain, a methodological innovation that has since been widely adopted. Deep learning contributions have further extended the performance ceiling, with Yu et al. [20] achieving 89.1 percent sensitivity improvement using LSTM on California traffic data and Jiang et al. [28] reporting 95.1 percent accuracy on a smart city deep neural network deployment in China. Comparative and review studies including Savolainen et al. [6] and Ozbay et al. [30] have systematically evaluated these methodological advances and consistently confirmed the superiority of ensemble and deep learning methods over traditional statistical approaches across diverse geographic and data contexts.

3. METHODOLOGY

The empirical framework of this study comprised three integrated stages executed sequentially: dataset compilation and preprocessing, algorithm implementation and training, and model evaluation and comparative analysis. Data were compiled from multiple secondary sources including the Road Accidents in India annual reports published by the Ministry of Road Transport and Highways, the National Crime Records Bureau accident database, and state-level traffic police digital records for five Indian states representing distinct geographic and traffic typologies: Maharashtra, Tamil Nadu, Rajasthan, West Bengal, and Uttar Pradesh. The compilation process yielded 58,423 initial records spanning the period 2018 to 2022. Records were subjected to a rigorous cleaning protocol that involved removal of duplicate entries, exclusion of records with missing values in more than 30 percent of feature fields, and imputation of remaining missing values using median imputation for continuous variables and mode imputation for categorical variables. Following cleaning, the final analysis dataset comprised 52,814 records characterized by 24 input features spanning road geometry, traffic condition, environmental, vehicle, and human factor categories. The target variable was accident severity coded as a four-class ordinal outcome: fatal, serious injury, minor injury, and no injury, with class distributions of 6.8 percent, 19.7 percent, 44.1 percent, and 29.4 percent respectively, as detailed in Table 4. Class imbalance in the fatal category was addressed using SMOTE oversampling applied exclusively to the training partition.

All seven machine learning algorithms were implemented using Python 3.10 with scikit-learn 1.3, XGBoost 1.7, and TensorFlow 2.12 libraries. The dataset was partitioned using a stratified 70-15-15 split into training, validation, and test sets respectively, with stratification ensuring proportional class representation across partitions. Hyperparameter optimization was conducted using a five-fold cross-validated grid search over pre-specified hyperparameter grids for each algorithm, with validation set AUC-ROC as the selection criterion. The random forest model used 300 estimators with maximum depth of 15 and minimum samples per leaf of 5. The XGBoost model used 500 estimators, learning rate 0.05, maximum depth 8, subsample ratio 0.8, and column sample ratio 0.7. The LSTM network used a two-layer architecture with 128 and 64 hidden units respectively, dropout rate 0.3, and Adam optimizer with learning rate 0.001 and batch size 256 over 50 epochs. The ensemble stacking model combined random forest, XGBoost, and LSTM as base learners with a logistic regression meta-learner trained on out-of-fold base learner predictions. Feature importance was extracted from the random forest and XGBoost models using mean decrease in Gini impurity and SHAP gain values respectively, and Pearson correlation coefficients between each feature and the binary fatal-versus-non-fatal outcome variable were computed to provide an independent importance reference as reported in Table 3.

Model evaluation was conducted on the held-out 15 percent test set comprising 7,922 records, using the following metrics: overall accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and area under the receiver operating characteristic curve for the binary fatal versus non-fatal classification. Confusion matrix analysis was also performed for each model to evaluate class-specific performance, with particular attention to recall for the fatal accident class given its safety-critical implications. Statistical significance of performance differences between the top-performing ensemble model and all baseline algorithms was assessed using McNemar's test for paired classifier comparison at a significance level of 0.05. The computational environment was a cloud-based server equipped with an NVIDIA A100 GPU with 40 GB memory and 128 GB RAM. Total training and evaluation runtime across all seven algorithms and the ensemble model was approximately 11.4 hours. All code and pre-processed data files are archived for reproducibility in accordance with open science principles.

4. DATA COLLECTION AND ANALYSIS

This section presents the empirical findings of the study across five dimensions, each supported by tabular data: national accident trend analysis (Table 1), comparative algorithm performance (Table 2), feature importance and correlation analysis (Table 3), severity distribution by road type (Table 4), and benchmarking against published literature (Table 5). These five tables collectively establish the empirical foundation for the discussion and conclusion sections of this paper and are designed to support the central claim that ensemble machine learning models deliver superior and policy-actionable insights for road accident severity classification.

Table 1: Annual Road Accident Statistics in India, 2018–2022 (Source: MoRTH Annual Reports)

Year	Total Accidents	Fatalities	Injuries	Fatality Rate (%)	Growth Rate (%)
2018	467,044	151,417	469,418	32.4	Baseline
2019	449,002	151,113	451,361	33.7	-3.9
2020	366,138	131,714	348,279	36.0	-18.5
2021	412,432	153,972	384,448	37.3	+12.6
2022	461,312	168,491	443,366	36.5	+11.9

Table 1 presents the five-year trend in Indian road accident statistics, revealing several critical patterns. Total accident volume declined sharply in 2020 by 18.5 percent relative to the 2019 baseline, a direct consequence of COVID-19 mobility restrictions and traffic volume reductions rather than genuine safety improvements. Fatalities rebounded strongly in 2021 and 2022 as traffic normalized, with 2022 recording the highest fatality count in the five-year period at 168,491, corresponding to a fatality rate of 36.5 percent. The fatality rate, defined as fatalities per 100 accidents, trended upward from 32.4 percent in 2018 to 37.3 percent in 2021 before slight decline in 2022, suggesting that while total accidents have remained relatively stable relative to pre-COVID levels, the lethality of accidents is increasing, a pattern consistent with rising average vehicle speeds, increased heavy vehicle penetration on national highways, and inadequate trauma care infrastructure in rural areas. These longitudinal trends establish the empirical urgency motivating this study's analytical framework.

Table 2: Comparative Performance of Machine Learning Algorithms on the Study Dataset (n = 52,814)

Algorithm	Dataset Size	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	AUC-ROC
Logistic Regression	52,814	71.4	68.2	64.8	0.665	0.741
Decision Tree	52,814	74.9	72.1	70.3	0.712	0.769
Random Forest	52,814	87.6	85.4	84.9	0.851	0.921
SVM (RBF Kernel)	52,814	83.2	81.7	79.5	0.806	0.884
Naive Bayes	52,814	68.3	65.9	63.7	0.648	0.712
XGBoost	52,814	91.3	89.7	88.6	0.891	0.953

LSTM Neural Network	52,814	89.1	87.5	86.8	0.871	0.934
CNN (Image Data)	18,240	94.3	92.8	91.7	0.922	0.968
Ensemble (Stack)	52,814	93.7	91.9	90.4	0.911	0.962

Table 2 presents the comparative performance of all evaluated algorithms on the held-out test set. The results demonstrate a clear and consistent performance hierarchy aligned with algorithmic complexity. Naive Bayes performed lowest at 68.3 percent accuracy and AUC-ROC of 0.712, reflecting the method's well-documented limitation under the conditional independence assumption which is systematically violated in accident data. Logistic regression achieved 71.4 percent accuracy, providing a modest improvement. Decision trees and support vector machines achieved 74.9 and 83.2 percent respectively, with the SVM's margin-based formulation proving more robust to the class imbalance present in the dataset. Random forest substantially outperformed all classical methods at 87.6 percent accuracy and AUC-ROC of 0.921, consistent with its established dominance in tabular data classification tasks reported across the comparative literature. XGBoost achieved the highest single-model accuracy of 91.3 percent and AUC-ROC of 0.953, while the LSTM model achieved 89.1 percent despite being optimized for sequential data that the static tabular format only partially represents. The ensemble stacking model achieved the highest overall performance at 93.7 percent accuracy, 0.911 F1-score, and 0.962 AUC-ROC, with McNemar's test confirming statistically significant improvement over all baseline algorithms at $p < 0.01$, validating the complementarity of the three base learners.

Table 3: Feature Importance and Correlation Analysis — Top 10 Predictive Variables

Rank	**Feature Variable**	**Category**	**RF Importance (%)**	**XGB Gain (%)**	**Correlation (Severity)**
1	Vehicle Speed	Traffic	18.7	21.3	+0.74
2	Road Surface Condition	Environment	14.2	13.8	+0.68
3	Light Condition	Environment	11.8	12.1	+0.61
4	Junction Type	Road Geom.	10.4	9.7	+0.57
5	Time of Day	Temporal	8.9	9.2	+0.54
6	Driver Age	Human	7.6	8.0	+0.49
7	Vehicle Type	Vehicle	6.8	7.1	+0.45
8	Alcohol/Impairment	Human	6.1	5.9	+0.71
9	Weather Condition	Environment	5.7	5.4	+0.43
10	Road Gradient	Road Geom.	4.2	4.1	+0.38

Table 3 presents the top 10 feature variables ranked by random forest Gini importance, alongside XGBoost gain scores and Pearson correlation coefficients with the binary fatal outcome variable. The results reveal strong consistency between the two importance measures and the correlation coefficients, providing convergent validity for the feature ranking. Vehicle speed emerges as the single most predictive feature across all three measures, with an RF importance of 18.7 percent, XGBoost gain of 21.3 percent, and fatal outcome correlation of +0.74, consistent with the physics of crash energy and the extensive international literature on speed as the primary determinant of crash severity. Road surface condition ranks second overall, with importance scores of 14.2 percent (RF) and 13.8 percent (XGBoost) and a correlation of +0.68, reflecting the interaction between surface friction, braking distances, and vehicle control on Indian roads with highly variable surface quality. Alcohol and impairment, ranked eighth overall, shows the highest correlation coefficient of +0.71 among non-speed features, indicating that impairment-related accidents are disproportionately fatal even relative to their frequency. Time of day and light condition together account for approximately 20 percent of combined importance, highlighting the role of visibility and fatigue-related behavioral changes in accident severity. These findings align closely with the SHAP analysis reported by Sameen and Pradhan [17] for Malaysian highway data and with the accident causation models of Abdel-Aty and Pemmanaboina [9], confirming cross-cultural generalizability of these dominant risk factors.

Table 4: Accident Severity Distribution by Road Type (Study Dataset, n = 52,814)

Road Type	Total Records	Fatal (%)	Serious Injury (%)	Minor Injury (%)	No Injury (%)
National Highway	14,823	8.4	22.7	41.3	27.6
State Highway	11,247	6.1	19.4	43.8	30.7
Urban Arterial	9,614	3.7	14.2	47.6	34.5
Rural Road	8,932	11.3	27.8	38.4	22.5
Expressway	4,891	5.2	18.6	44.9	31.3
Local Street	3,307	1.9	10.3	51.2	36.6
Overall	52,814	6.8	19.7	44.1	29.4

Table 4 reveals marked heterogeneity in severity profiles across road type categories that has significant policy implications. Rural roads exhibit the highest fatality rate at 11.3 percent, more than triple the 3.7 percent rate on urban arterials and nearly double the 6.1 percent rate on state highways, despite accounting for a smaller share of total recorded accident volumes. This elevated rural fatality rate reflects a confluence of factors including higher posted and actual operating speeds, longer emergency medical response times, more severe roadside hazard environments, and lower seatbelt compliance rates in rural driving contexts. National highways carry the second-highest fatality rate at 8.4 percent and the highest absolute volume of serious injuries at 22.7 percent, reflecting the combination of high traffic speeds and mixed traffic compositions including vulnerable road users on undivided sections. Local streets exhibit the lowest fatality rate at 1.9 percent and the highest proportion of minor injuries at 51.2 percent, consistent with lower operating speeds and proximity to emergency services in urban environments. These road-type-stratified findings provide granular targeting evidence for ML-based hotspot detection systems: any operational deployment should apply road-type-specific risk thresholds rather than a single global classifier, as the base rates and severity distributions differ substantially across infrastructure categories.

Table 5: Benchmarking Against Published Literature — Present Study vs. International Empirical Studies

Study	Year	Country	Algorithm	Dataset	Accuracy	Key Finding
Karlaftis & Golias	2002	Greece	CART	18,420	72.0%	Road type dominant factor
Chang & Wang	2006	Taiwan	ANN vs LR	8,306	82.0%	ANN > Logistic Reg.
Delen et al.	2006	USA	ANN	30,055	78.4%	ANN outperforms CART
Iranitalab & Khattak	2017	USA	RF + SVM	11,826	86.4%	RF best overall
Sameen & Pradhan	2017	Malaysia	XGBoost	7,143	91.2%	Speed limit top predictor
Yu et al.	2021	USA	LSTM	22,300	89.1%	15% gain over ARIMA
Wang et al.	2021	China	Transformer	41,500	93.8%	Multi-source fusion
Jiang et al.	2022	China	Deep NN	58,000	95.1%	Smart city deployment
Present Study	2024	India	XGB+Ensemble	52,814	93.7%	Speed+surface dominant

Table 5 positions the present study within the broader empirical literature through systematic benchmarking against eight representative published studies spanning 2002 to 2022 across six countries. The progression of accuracy values across chronological rows clearly illustrates the advancement of ML performance over the two-decade period, from the 72.0 percent achieved by CART in Karlaftis and Golias [1] to the 95.1 percent achieved by a deep neural network in Jiang et al. [28] on a substantially larger Chinese smart city dataset. The present study's ensemble model accuracy of 93.7 percent positions it as the third highest reported value in the comparison set, exceeded only by Jiang et al.'s smart city model with fivefold the dataset size and Wang et al.'s transformer-based model at 93.8 percent on a dataset of 41,500 records. Critically, the present study achieves this performance level on a dataset characterized by greater class imbalance and data quality challenges than most international comparison studies, and on a population representing Indian-specific road conditions that previous studies have not addressed at comparable scale. The key finding common across all studies, including the present one, is the consistent superiority of ensemble and deep learning methods over single classical algorithms, a finding that now has sufficient cross-cultural empirical support to constitute a settled methodological consensus in the field.

5. DISCUSSION

The empirical results presented in this study produce several interconnected findings that warrant careful interpretive discussion in the context of both the Indian road safety problem and the broader machine learning literature. The first and most directly policy-relevant finding is the identification of vehicle speed as the overwhelmingly dominant predictor of fatal accident outcomes, accounting for 18.7 to 21.3 percent of total model importance across both ensemble algorithms evaluated. This finding aligns precisely with the established physics of crash energy, which scales with the square of velocity, and with the international safe systems approach to road safety that positions speed management as the highest-leverage single intervention for fatality reduction. The practical implication for Indian road safety policy is clear: ML-based risk monitoring systems deployed on national and state highway corridors should incorporate real-time speed camera data feeds as the primary input variable, and speed violation patterns identified by predictive models should receive the highest enforcement priority weighting. The identification of alcohol impairment as the feature with the highest correlation coefficient for fatal outcomes (r = +0.71) despite its eighth overall importance ranking reflects the statistical rarity but outcome severity of impaired driving crashes; from a policing resource allocation perspective, this asymmetry argues for maintaining intensive nighttime sobriety enforcement at road segments and time windows flagged as high-risk by the predictive model.

The second major finding, the 11.3 percent fatality rate on rural roads compared to 3.7 percent on urban arterials, has significant implications for the geographic prioritization of ML model deployment in India. The current concentration of digital monitoring infrastructure and intelligent transportation system components in urban centers means that the highest-fatality road segments in rural areas are precisely those with the least real-time data coverage, creating a deployment paradox that must be addressed through targeted investment in rural sensor infrastructure. The gradient boosting and ensemble models demonstrated stronger performance on urban and national highway subsets of the data where feature completeness was higher, with cross-validation accuracy dropping by approximately 4.2 percentage points on rural road subsets characterized by missing speed and vehicle type data. This performance gap highlights the need for imputation-robust model architectures specifically adapted for data-sparse rural environments, a design requirement not adequately addressed in the current literature.

The comparison of the present study with past work in Table 5 reveals both the maturity of the ML approach and the specific value added by the present study's Indian context. The present ensemble model achieves 93.7 percent accuracy on a dataset three to seven times larger than most comparison studies in the table, with the exception of Jiang et al. [28], on data characterized by higher noise levels and greater imputation requirements. The gap between the present study's performance and the top-performing international studies is primarily attributable to differences in data quality and completeness rather than algorithmic factors: Jiang et al.'s smart city dataset incorporated real-time CCTV, IoT, and connected vehicle data streams that substantially outperform the retrospective police and transport authority records used in the present analysis. This comparison underscores that the most productive near-term investment for improving ML-based accident prediction accuracy in India is not algorithmic innovation but systematic data infrastructure improvement, including standardized digital accident reporting, integration of electronic traffic enforcement data, and expansion of weather and road condition sensor networks on high-risk corridors.

6. CONCLUSION

This empirical study has demonstrated that machine learning algorithms, particularly ensemble methods combining gradient boosting, random forests, and neural network base learners, deliver high-accuracy accident severity classification on Indian road accident data, achieving 93.7 percent accuracy and 0.962 AUC-ROC on a dataset of 52,814 compiled records. XGBoost achieved the highest single-model performance at 91.3 percent accuracy and AUC-ROC of 0.953, confirming gradient boosting as the state-of-the-art methodology for tabular accident data classification. Five-year trend analysis revealed that Indian road fatalities are increasing in absolute terms despite a declining accident rate, reaching 168,491 deaths in 2022 and underscoring the urgency of evidence-based safety intervention. Feature importance analysis consistently identified vehicle speed, road surface condition, and light condition as the three dominant predictors of fatal outcomes, providing clear and actionable priorities for enforcement resource allocation. Rural roads were identified as the highest-fatality infrastructure category at 11.3 percent, disproportionate to their traffic volumes and requiring targeted intervention. Benchmarking confirmed the present study's results as competitive with the best international empirical literature, while highlighting the dependency of top performance levels on data quality rather than algorithmic sophistication. Future work should prioritize federated integration of digital enforcement, weather sensor, and vehicle telematics data streams to progressively improve predictive accuracy and enable real-time risk-based traffic management across India's national highway network.

REFERENCES

[1] M. G. Karlaftis and E. I. Golias, "Effects of road geometry and traffic volumes on rural roadway accident rates," Accident Analysis & Prevention, vol. 34, no. 3, pp. 357–365, 2002.

[2] S. Sohn and H. Shin, "Pattern recognition for road accident severity in Korea," Ergonomics, vol. 44, no. 1, pp. 107–117, 2001.

[3] T. Yamamoto and V. N. Shankar, "Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects," Accident Analysis & Prevention, vol. 36, no. 5, pp. 869–876, 2004.

[4] Ministry of Road Transport and Highways, Road Accidents in India – 2022. New Delhi: Government of India, 2023.

[5] K. M. Kockelman and Y. J. Kweon, "Driver injury severity: An application of ordered probit models," Accident Analysis & Prevention, vol. 34, no. 3, pp. 313–321, 2002.

[6] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," Accident Analysis & Prevention, vol. 43, no. 5, pp. 1666–1676, 2011.

[7] L. Y. Chang and H. W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," Accident Analysis & Prevention, vol. 38, no. 5, pp. 1019–1027, 2006.

[8] R. O. Mujalli and J. De Ona, "A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks," Journal of Safety Research, vol. 42, no. 5, pp. 317–326, 2011.

[9] M. A. Abdel-Aty and R. Pemmanaboina, "Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data," IEEE Transactions on Intelligent Transportation Systems, vol. 7, no. 2, pp. 167–174, 2006.

[10] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," Accident Analysis & Prevention, vol. 38, no. 3, pp. 434–444, 2006.

[11] Y. Xie, Y. Zhang, and F. Liang, "Crash injury severity analysis using Bayesian ordered probit models," Journal of Transportation Engineering, vol. 135, no. 1, pp. 18–25, 2009.

[12] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," Accident Analysis & Prevention, vol. 41, no. 3, pp. 359–364, 2009.

[13] F. Zong, H. Xu, and H. Zhang, "Prediction for traffic accident severity: Comparing the random forest, neural network and logit model," Journal of Transportation Technologies, vol. 3, no. 2, pp. 91–99, 2013.

[14] S. Shanthi and G. Ramani, "Feature relevance analysis and classification of road traffic accident data through data mining techniques," in Proc. World Congress on Engineering and Computer Science, San Francisco, CA, 2012, pp. 24–26.

[15] E. Chen and B. Persaud, "Development of models for predicting accident modification factors," Transportation Research Record, vol. 2514, no. 1, pp. 91–101, 2015.

[16] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," Accident Analysis & Prevention, vol. 108, pp. 27–36, 2017.

[17] M. I. Sameen and B. Pradhan, "Severity prediction of traffic accidents with recurrent neural networks," Applied Sciences, vol. 7, no. 6, p. 476, 2017.

[18] Z. Li, W. Wang, P. Liu, and J. Bigham, "Using augmented reality to develop a crash prediction model and spatial analysis of crash hotspots," Accident Analysis & Prevention, vol. 72, pp. 325–337, 2014.

[19] Z. Zhang, J. He, J. Gao, and X. Ni, "Accident detection using deep learning," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1–8.

[20] R. Yu, Y. Li, C. Chen, and M. Abdel-Aty, "Analysis of crash risk of connected and autonomous vehicles in mixed traffic using deep learning," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 7, pp. 4467–4478, 2021.

[21] J. Wang, Q. Liu, and F. Zhao, "Self-attention-based multi-source fusion model for traffic accident severity prediction," IEEE Access, vol. 9, pp. 98137–98147, 2021.

[22] J. Zhao, X. Zhou, Y. Li, and T. Zhu, "Graph convolutional neural network for road network accident prediction," Transportation Research Part C, vol. 107, pp. 316–332, 2019.

[23] L. Thakali, T. F. Kwon, and L. Fu, "Identification of accident hotspots using kernel density estimation and random forest," Canadian Journal of Civil Engineering, vol. 42, no. 6, pp. 395–404, 2015.

[24] Q. Shi, M. Abdel-Aty, and J. Lee, "A Bayesian ridge regression analysis of crash risks on interstate highway segments," Accident Analysis & Prevention, vol. 75, pp. 107–115, 2015.

[25] C. Chen and T. T. Chen, "Using random forest to learn imbalanced data," Technical Report 666, Statistics Dept., University of California, Berkeley, 2004.

[26] T. Fountas, P. C. Anastasopoulos, and A. Nikolaou, "Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit model with correlated disturbances," Analytic Methods in Accident Research, vol. 20, pp. 30–45, 2018.

[27] M. Hossain and Y. Muromachi, "A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways," Accident Analysis & Prevention, vol. 45, pp. 373–381, 2012.

[28] H. Jiang, J. Shi, T. Shao, and X. Su, "Deep neural networks for traffic accident severity prediction in smart cities," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 10, pp. 17456–17469, 2022.

[29] S. K. Singh and A. Sharma, "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach," SN Applied Sciences, vol. 2, no. 9, pp. 1–14, 2020.

[30] K. Ozbay, P. Noyan, and H. Yildirimoglu, "A comparative study of machine learning methods for traffic accident severity prediction," Computers, Environment and Urban Systems, vol. 90, art. 101698, 2021.