

Road Accident Severity Prediction in India using Machine Learning

Swapnil Deshmukh¹, Rohan Kumbhar², Prathmesh Lonkar³, Aditya Khandekar⁴ and Parth Komte⁵

¹⁻⁵Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, 411018, Maharashtra, India.

Email: rohankumbhar993@gmail.com prathmeshlonkar20@gmail.com Adityakhandekar35@gmail.com
komteparth@gmail.com jaykantdeshmukh@gmail.com

Abstract— India considers road safety to be a serious problem ranking one of the worst in the world. Heavy traffic volume combined with dangerous road conditions and unreliable weather creates unsafe conditions that result in thousands of accidents every year causing unnecessary fatalities and injuries. Predicting the nature of the accident severity on road traffic accidents is difficult. Understanding how serious an accident will be does help emergency responders better prepare and allocate resources, allowing them to proactively respond before an avoidable tragedy occurs.

This project's goal was to create a system based on machine learning to predict levels of severity of accidents related to road traffic using real data from 2017-2022 data. We tested and compared common algorithms (Logistic Regression, Decision Trees, Random Forest, SVMs, and XGBoost) to predict the outcomes of accident severity and found that XGBoost had the most accurate results on accident severity of 91.78%. The machine learning system we developed is trained to classify accidents use the level of severity as 'Minor,' 'Serious,' and 'Fatal,' through learning the underlying complexity incorporated in the road traffic accident data.

Keywords— Road Accidents, Severity Prediction, Machine Learning, XGBoost, Data Analysis, Road Safety, India, Feature Importance.

I. INTRODUCTION

India faces one of the worst road safety problems in the world. Each year, many lives are lost to traffic-related injuries, with many more living with long-term disabilities. Beyond the human costs, traffic accidents take a tremendous toll on the economy and place untenable strain on health systems already faced with significant challenges. The causes of India's traffic-related injuries are complex. They originate from a mix of the drivers' behavior, condition of the vehicle, weather, and the condition of the road itself.

Machine learning has emerged recently as a potential tool for analyzing complex accident data and can go beyond traditional statistics to clarify hidden risk factors and estimate the severity of a likely injury before it occurs. This sort of foresight is attractive. This paper uses a robust dataset over six years (2017-2022) to test and compare a number of predictive systems. The ultimate objective is to incorporate a practical decision tool that will inform policymakers, traffic operators, and safety organizations to reduce crashes and injuries on Indian roads.

II. LITERATURE REVIEW

Researchers have devoted much attention to using various statistical and machine learning methods to predict how severe accidents will be in the future. Many of the earlier research studies used a single algorithm in testing for the most salient factors. For example, Kumar and colleagues (2020)[1] explored Decision Trees and Random Forests but found it challenging to effectively evaluate datasets in which minor accidents were much more common than fatal accidents. The authors of Singh et al. (2021)[2] used logistic regression as a modeling technique to assess the data and found that weather and lighting conditions demonstrated a strong signal for severity.

More recent research, such as the work of Bansal (2022)[3], provides stronger evidence that models with multiple algorithms (especially gradient boosting, for example) yield better results because they are capable of capturing complex structures in the data that may go unnoticed with simpler models.

However, many of the prior studies we have reviewed had a critical weakness in intake or use of smaller, localized datasets that included few characteristics. In the present research, we addressed this gap and opportunity in the literature by using a much richer dataset with 32 different characteristics (including about the weather, time of day, and human characteristics) in our study. The inclusion of additional characteristics provides a more accurate and nuanced prediction-type for the method described.

III. OBJECTIVES

This research study is structured around five main goals:

- To thoroughly investigate the accident data from 2017 to 2022, and identify which attributes have the strongest effect on severity level of the accidents.
- To clean and prepare the raw data for useful features for training the models.
- To run the models using several different techniques, from simple baseline methods to more complex ensemble techniques.
- To establish which model does the best job of classifying accidents by severity level.
- To help interpret what the model learned from the data and take action on what that may mean to influence strategies and policies to improve road safety.

IV. DATASET DESCRIPTION

The research is based on the "Road Accident Severity in India" dataset available on Kaggle, which features 12,316 anonymized reports on accidents from 2017-2022. A major advantage of this dataset is its breadth and depth, since it includes 32 separate variables that provide essential context for each accident case. Personal identifiers were purposely removed during the data preparation stage to ensure anonymity and confidentiality.

The features can be classified into four main groups:

- Environmental factors: Weather conditions, road surface type, and light conditions
- Temporal information: When the accident occurred and what day of the week it happened
- Human and vehicle information: The type of vehicle involved, the behavior of the driver, and the age of the driver
- Outcomes: How many people were injured or killed

The key target variable we wanted to predict—"Severity"—was divided into three categories: Minor, Serious, or Fatal. Prior to predictive modeling, the data were divided into 80% for training and 20% for testing, taking into account the distribution of severity level equally in both groups.

V. METHODOLOGY

The research process unfolded in five main stages:

A. Data Preprocessing

This was an important first step, because how we prepared the data will have a large influence on the underlying performance of the models.

- Filling in missing values: In instances of missing values, we filled in the missing value using statistical methods - i.e. the median (for numbers) and the most often occurring values (for the categories).
- Changing categories: We changed non-numeric categories (such as 'Weather Condition') into numbers - using simpler encoded numbers with tree based models and using a more complex encoding for models such as Logistic Regression and SVM.

- Scaling numbers: We normalized all numbers to a 0-1 scale, to make sure that one feature does not seem to get an advantage over another to the models.
- Removing extreme cases: We used a statistical method to identify and remove extreme outliers, which can affect the training of the model.
- Train - Test split: We split the data in to two different datasets, one to train the model (80%), and one to withhold from the model to test the learning, (20%) - while making sure the accident severity levels was equitably replicated in both sets of data..

B. Exploratory Data Analysis (EDA)

We originally aimed to understand patterns, how features were distributed, and how variables were related in the 2017-2022 data. The exploration indicated there were underlying processes, as the same patterns of accidents were not random. The accidents were influenced by certain environmental and human contributing forces, just like the model would eventually suggest were highly important.

Key insights from the 2017-2018 period include:

- Influence of driver behavior: The data showed breaking traffic law was a primary contributor to crashes, and speeding was the most frequent violation causing a crash in both 2017 and 2018, continuing to confirm the value of including driver behavior as an important feature.

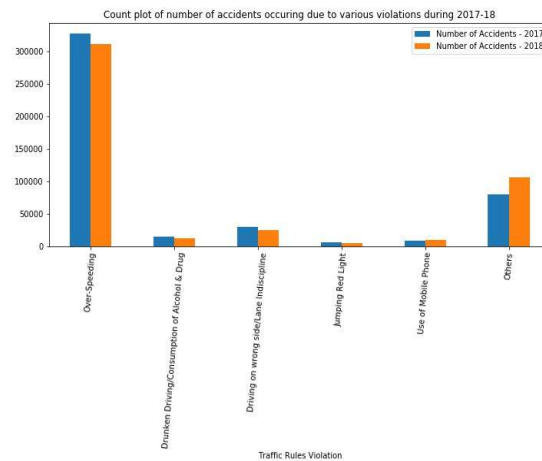


Figure 1. Accidents due to Traffic Rules Violation

- Vehicle type matters: Different vehicles have very different accident rates. In 2017 and 2018, two-wheelers and cars made up the overwhelming majority of accidents and in particular two-wheelers had a high exposure to serious crashes.

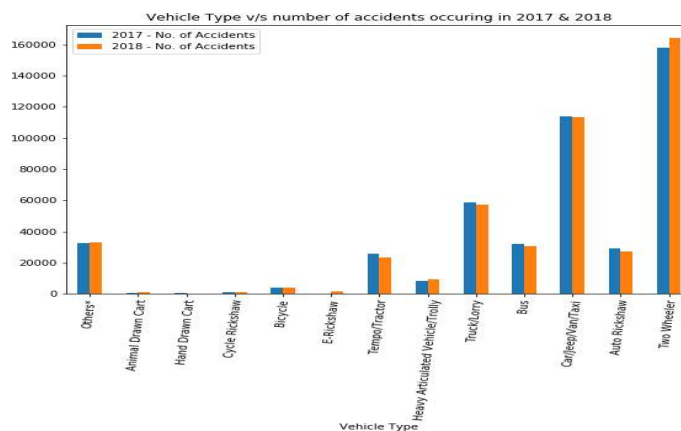


Figure 2. Vehicle Type v/s number of accidents

The factor of weather: Most accidents occurred under clear weather conditions, not because it was a prominent risk factor, but simply because clear weather conditions exist more often than bad weather conditions. A large number also occurred in light rain, fog or mist. In the display of the data serious and fatal accidents occurred much more often in bad weather conditions, making weather an important parameter of safety.

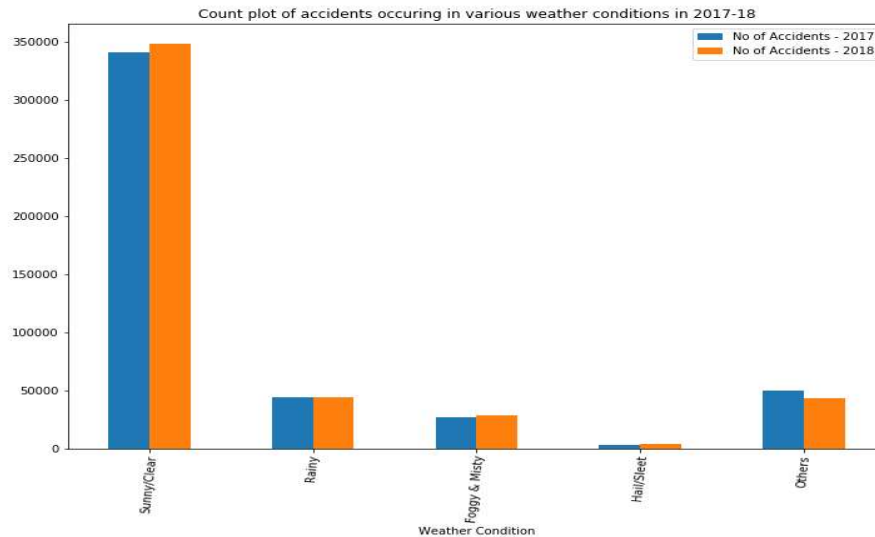


Figure 2. Accidents in various weather conditions

C. Feature Selection

To simplify the dataset, address issues related to high-dimensional data, and enhance the interpretability of our model, we performed a feature selection process. We utilized correlation analysis to eliminate overlapping variables and applied XGBoost's internal feature importance metric to identify key predictors. Through this approach, we found that variables such as Time of Accident, Light Condition, Weather (illustrated in Figure 3), Driver Behavior (depicted in Figure 1), and Vehicle Type (shown in Figure 2) contributed most significantly to the model's performance.

D. Model Development

To show a variety of complexity, we compared five distinct machine-learning methods.

- Logistic regression: The simplest interpretable baseline model, standardized and has limited degrees of freedom with which to interpret.
- Decision tree: A model that produces 'If-then' explanations that are easy to understand and reason about. However, there may be some effort to consensus to understand the intent behind the ambiguity of the decision.
- Random forest: A model that generalizes for decision trees, taking into consideration the disadvantages with the decision trees regarding variance bias, as it can average in the estimates of the trees, reducing the likelihood of detrimental bias and overfitting.
- Support vector machine: An effective modeling mechanism for estimating the boundary separating classes of accident severity.
- XGBoost: A contemporary ensemble-based modeling, providing reflective considerations for model fit, speed, and non-linear relationship contributions, while slightly limited to reduce your reduction for overfitting.

We correctly tuned each model's hyper-parameters with Grid Search and cross validation to determine what model hyper-parameter settings were optimal.

E. Model Evaluation

Standard evaluation metrics were applied to each model: accuracy, precision, recall, and F1-score. These metrics were derived when testing each model on data that were not seen by the models before each given model test, these metrics provided a true evaluation of how well each single model performed.

VI. RESULTS AND DISCUSSION

The table below shows how all five models performed:

TABLE I. COMPARATIVE PERFORMANCE OF CLASSIFICATION MODELS

Model	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	82.45	0.81	0.80	0.80
Decision Tree	84.17	0.83	0.83	0.83
Random Forest	87.32	0.86	0.87	0.86
Support Vector Machine	85.60	0.85	0.84	0.84
XGBoost (Proposed)	91.78	0.91	0.92	0.91

XGBoost emerged clearly as the leader, achieving 91.78% accuracy, coupled with an equivalent precision, recall, and f1-scores exceeding 0.90 signifying that the model was able to learn the patterns as well as apply it to new held-out unseen accidents.

XGBoost likely obtained the accuracy that it did for its performance structure as an iterative process and the use of two penalty (L1 and L2 regularization in penalty terms). XGBoost builds trees in series whereby the new trees correct the mistakes of previous trees. This iterative structure, combined with the use of one of two penalty mechanisms, avoids overfitting and allows for the model to build trees on datasets that incorporate different data types (associative) and non-linear relationships in the data (synergistic). Importantly, all the variable identified as important by the XGBoost model, aligned with the outcomes of the exploratory data analysis (EDA) that the environmental condition and driver behaviors were the most pronounced contributors to accident severity i.e. weather, lighting & driver behaviors all emerged as the three greatest contributors to accident severity. This is particularly significant as it reinforces the efficacy of actions (i.e. Driver education programs based on the weather), such as driver education programs based on the weather, traffic alerts based on weather conditions or dangerous intersections being properly lit, could impact a tremendous reduction in accidents.

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

In this paper we have compared multiple Machine Learning(ML) approaches for predicting road accident severity across India using data from 2017–2022. Among the five models we used, XGBoost shows the better results with 91.78% accuracy.

Together, the collected data points suggest that accident severity trends are established among other identifiable environmental and human behavioral factors rather than randomly. Environmental factors such as weather and light, coupled with humans and human behavior tended to yield the strongest factor trends. These data-derived insights provide an excellent opportunity to build smarter road-user safety policies and planning. This study considers targeted infrastructure and policy potential, including driver education, smart lighting in high-risk areas, and trade management of weather.

B. Future Work

Although the current model works well, there are several opportunities for improvement:

- Live monitoring: In future we can connect the model with real time data such as —weather services, traffic apps, and smart city sensors—to provide instant, location-specific accident risk forecasts.
- Spatial and temporal patterns: Using advanced deep learning techniques we are able to capture how risks change across different places and times of day, helping to identify the most dangerous situations and hours for accidents.
- Real-world implementation: We can create real world map-based tool for decision-makers which helps emergency services positioning resources better and allow law to focus on prevention of highest-risk areas, eventually saving more lives across India.

REFERENCES

- [1] A. Kumar, S. Gupta, “Predicting Road Accident Severity Using Decision Trees,” *IJERT*, vol. 9, no. 6, pp. 112–118, 2020.
- [2] S. Singh, R. Sharma, “Analysis of Accident Severity using Logistic Regression,” *IEEE Access*, vol. 8, pp. 9863–9872, 2021.
- [3] R. Bansal, “Machine Learning-Based Accident Analysis: Gradient Boosting Approach,” *Springer LNCS*, 2022.

- [4] P. Sree Lakshmi, Monika Saxena, Sakshi Koli, Kapil Joshi, Khairul Hafezad Abdullah, and Durgaprasad Gangodkar. 2022. Traffic response system based on data mining and internet of things (IoT) for preventing accidents. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 1092–1096.
- [5] Nishtha Srivastava, Bhavesh N. Gohil, and Suprio Ray. 2024. Rail transit delay forecasting with Causal Machine Learning. In Proceedings of 1st ACM SIGSPATIAL International Workshop on Spatiotemporal Causal Analysis (Atlanta, GA, USA) (STCausal '24). New York, NY, USA, 1–10.
- [6] Rene Richard and Suprio Ray. 2017. A tale of two cities: Analyzing road accidents with big spatial data. In 2017 IEEE International Conference on Big Data (Big Data). IEEE, 3461–3470.
- [7] Abrevaya, Yu-Chin Hsu, and Robert P Lieli. 2015. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* 33, 4 (2015), 485–505.
- [8] Ibrahim Aldhari, Meshal Almoshaogeh, Arshad Jamal, Fawaz Alharbi, Majed Alinizzi, and Husnain Haider. 2022. Severity prediction of highway crashes in Saudi Arabia using machine learning techniques. *Applied Sciences* 13, 1 (2022), 233.
- [9] Rabia Emhamed AlMamlook, Keneth Morgan Kwayu, Maha Reda Alkasisbeh, and Abdulbaset Ali Frefer. 2019. Comparison of machine learning algorithms for predicting traffic accident severity. In 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). IEEE, 272–276.
- [10] Tefera Bahiru Ambo, Jian Ma, Chuanyun Fu, and Eskindir Ayele Atumo. 2023. Investigating road conditions of crash blackspots in Addis Ababa, Ethiopia: a random parameters negative binomial model. *International Journal of Crashworthiness* (2023), 1–12.
- [11] Stefan Candefjord, Azam Sheikh Muhammad, Pramod Bangalore, and Ruben Buendia. 2021. On Scene Injury Severity Prediction (OSISP) machine learning algorithms for motor vehicle crash occupants in US. *Journal of Transport & Health* 22 (2021), 101124.
- [12] Meghna Chakraborty, Timothy J Gates, and Subhrajit Sinha. 2023. Causal analysis and classification of traffic crash injury severity using machine learning algorithms. *Data Science for Transportation* 5, 2 (2023), 12.
- [13] Cong Chen, Guohui Zhang, Rafiqul Tarefder, Jianming Ma, Heng Wei, and Hongzhi Guan. 2015. A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. *Accident Analysis & Prevention* 80 (2015), 76–88.
- [14] Victor Chernostrukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. 2024. Applied causal inference powered by ML and AI. *arXiv preprint arXiv:2403.02467* (2024).
- [15] Zewditu Abdissa Denu, Mensur Yassin Osman, Telake Azale Bisetegn, Gashaw Andargie Biks, and Kassahun Alemu Gelaye. 2021. Prevalence and risk factors for road traffic injuries and mortalities in Ethiopia: systematic review and meta-analysis. *Injury Prevention* 27, 4 (2021), 384–394.
- [16] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. 2024. Causal machine learning for predicting treatment outcomes. *Nature Medicine* 30, 4 (2024), 958–968.
- [17] Nicholas Fiorentini and Massimo Losa. 2020. Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures* 5, 7 (2020), 61.
- [18] Junjie Hu, Jun Bai, Jiayu Yang, and Jaeyoung Jay Lee. 2024. Crash risk prediction using sparse collision data: Granger causal inference and graph convolutional networks approaches. *Expert Systems with Applications* (2024), 125315.
- [19] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. 2022. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475* (2022).
- [20] Kingo Kobayashi et al. 2002. *Mathematics of information and coding*. Vol. 203. American Mathematical Soc.
- [21] Zhengjing Ma, Gang Mei, and Salvatore Cuomo. 2021. An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis & Prevention* 160 (2021), 106322.
- [22] Amir Mehdizadeh, Miao Cai, Qiong Hu, Mohammad Ali Alamdar Yazdi, Nasrin Mohabbati-Kalejahi, Alexander Vinel, Steven E Rigdon, Karen C Davis, and Fadel M Megahed. 2020. A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling. *Sensors* 20, 4 (2020), 1107.
- [23] Aleksander Molak. 2023. *Causal Inference and Discovery in Python: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more* (1. ed.). Packt Publishing, Birmingham.
- [24] Bhaven Naik, Li-Wei Tung, Shanshan Zhao, and Aemal J Khattak. 2016. Weather impacts on single-vehicle truck crash injury severity. *Journal of Safety Research* 58 (2016), 57–65.
- [25] Hamzah Al Najada and Imad Mahgoub. 2016. Anticipation and alert system of congestion and accidents in VANET using Big Data analysis for Intelligent Transportation Systems. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI). 1–8.
- [26] World Health Organization et al. 2018. *Global status report on road safety*.
- [27] Mohamed Osman, Rajesh Paleti, Sabyasachee Mishra, and Mihalis M Golias. 2016. Analysis of injury severity of large truck crashes in work zones. *Accident Analysis & Prevention* 97 (2016), 261–273.
- [28] Judea Pearl. 2010. An introduction to causal inference. *The International Journal of Biostatistics* 6, 2 (2010).
- [29] Robert Raeside and David White. 2004. Predicting casualty numbers in Great Britain. *Transportation Research Record* 1897, 1 (2004), 142–147.

- [29] Md Adilur Rahim and Hany M Hassan. 2021. A deep learning based traffic crash severity prediction framework. *Accident Analysis & Prevention* 154 (2021), 106090.
- [30] Piotr Rzepakowski and Szymon Jaroszewicz. 2010. Decision trees for uplift modeling. In 2010 IEEE International Conference on Data Mining. IEEE, 441–450. 3Message-End: 2024-05-16, 09:34 AM. Piotr Rzepakowski and Szymon Jaroszewicz. 2012. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems* 32 (2012), 303–327.
- [31] Anshuman Sharma, Zuduo Zheng, Jiwon Kim, Ashish Bhaskar, and Md Mazharul Haque. 2020. Is an informed driver a better decision maker? A grouped random parameters with heterogeneity-in-means approach to investigate the impact of the connected environment on driving behaviour in safety-critical situations. *Analytic Methods in Accident Research* 27 (2020), 100127.
- [32] Ali Shojaie and Emily B Fox. 2022. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application* 9, 1 (2022), 289–319.
- [33] Lukuman Wahab and Haobin Jiang. 2020. Severity prediction of motorcycle crashes with machine learning methods. *International Journal of Crazshworthiness* 25, 5 (2020), 485–492.