Literature Review on Crash Reporting - Incidents Data

Sai Anish Sreeramagiri

March 2024

1 Introduction

Road Traffic Accidents continue to be a significant concern all over the world, resulting in loss of life, property damage, injuries, and economic burden. Road traffic accidents are a global crisis, claiming the lives of 1.3 million people and causing nearly 50 million non-fatal injuries annually. To mitigate these incidents and improve road safety, it is crucial to identify their root causes. So, traffic crash data analysis plays a pivotal role in understanding the dynamics of road safety, assisting policymakers and law enforcement agencies in implementing effective measures to reduce accidents and their impacts. Such analyses rely heavily on comprehensive datasets that capture various factors surrounding each incident. This literature review delves into the surveying of various works done on traffic crash data and using it to solve different problems with the help of various machine learning algorithms. These papers also discuss various parameters highlighting their significance in clarifying the circumstances around accidents.

The datasets include lots of crucial information such as report numbers, weather conditions, driver behavior, surface conditions, vehicle characteristics, and injury severities. These characteristics give important information about the conditions surrounding the accidents to recognize the underlying causes and potential areas for action. Various problems can be solved using the traffic accident dataset such as crash severity prediction, accident risk prediction, accident-prone hotspots, etc. With the help of various machine learning algorithms and different feature selection strategies, we can understand the different root causes behind accidents, which are discussed in the below papers.

2 Literature Review

2.1 Comprehensive Analysis of Road Traffic Accidents in the United States: Unveiling Trends, Patterns, Predictions, and Road Safety Insights through Big Data Analysis[Aas24]

It is an in-depth study conducted to understand various underlying factors behind road traffic accidents in Montgomery, United States. Here, they identified many patterns, correlations, important features, and trends related to traffic accidents. The analysis revealed critical insights into the factors contributing to road traffic accidents, like environmental conditions, driver behavior, and vehicle characteristics. Here, the "Driver at Fault" column is a potential target variable for classification tasks, indicating whether the driver was at fault in the collision. After conducting these various visualizations and analyses, they identified multiple key features affecting this target variable. They used the Random Forest Classifier which gave an accuracy of 86% with a 0.92 AUC score and the XGBoost Classifier with an accuracy of 87% with a 0.94 AUC score. [Aas24]

2.2 Automatic Classification of Traffic Incident's Severity Using Machine Learning Approaches[Hoa17]

This paper discusses a real-time automatic decision support system that can minimize a traffic operator's response time and reduce congestion. Apart from standard descriptions (incident/crash location, time, date, and lane affected), severity is an important criterion that operators have to evaluate based on all available information. Data has been collected from Sydney's Transport Management Center and was used to visualize frequent patterns in historical incident response records, leading to the automatic classification of accident severity using advanced machine learning, active learning, and outlier detection techniques. Using 40,000 records from the past data, the classification model achieved nearly 90% accuracy in five-fold cross-validation. Algorithms used were Naive Bayes, KNNs, SVMs, and Decision Trees.[Hoa17]

2.3 Data Visualization and Machine Learning Approach for Analyzing Severity of Road Accidents[Ras22]

In this study, the authors performed an in-depth descriptive analysis to identify significant factors influencing accident severity. Considering factors like location, time, infrastructure, and environmental conditions that may cause road accidents, they have leveraged machine learning techniques to predict the severity of accidents. According to their study, the frequency of medium-risk accidents is very high compared to low and high-risk accidents. Moreover, factors like infrastructure, day of the week, and weather conditions influence the severity of accidents differently. The authors implemented Logistic Regression, KNN, Decision Tree, and Random Forest algorithm in predicting the severity of road traffic accidents where the Random Forest algorithm gave an accuracy of 97.2%.[Ras22]

2.4 Crash severity analysis of vulnerable road users using machine learning[Kom21]

This paper focuses on employing machine learning-based classification approaches for modeling the injury severity of vulnerable road users(VRU) namely pedestrians, bicyclists, and motorcyclists. The critical factor of crash severity outcomes for these VRU groups is estimated to identify the similarities and differences across different important features associated with different VRU groups. Here, 17 distinct road crash parameters are considered as input features to train models which range from road user characteristics, weather and environment, vehicle and driver condition, period, road characteristics and regions, traffic, and speed jurisdiction. Classification models such as KNNs, SVMs, and Random Forest are separately trained and tested for individual and unified VRU to assess crash severity levels. Model performances are compared with each other to justify the best classifier where Random Forest classification models for all VRU modes are found to be comparatively robust in test accuracy: (motorcyclist: 72.30%, bicyclist: 64.45%, pedestrian: 67.23%, unified VRU: 68.57%). Furthermore, a model-based partial dependency of each road crash parameter on the severity levels is plotted and compared for each individual and unified VRU. Based on the comparative analysis's outcome, motorcyclists are more likely exposed to higher crash severity, followed by pedestrians and bicyclists.[Kom21]

2.5 Comparative analysis of pedestrian crash severity at United Kingdom rural road intersections and Non-Intersections using latent class clustering and ordered probit model[Ali23]

This study aimed to investigate the severity of pedestrian injuries sustained on rural roads in the UK, including crashes at intersections and non-intersections. Here, the authors utilized Latent Class Analysis to identify homogenous clusters of clashes. Additionally, they utilized the Ordered Probit Model to identify contributing factors within each cluster. Factors like pedestrian location in footways and one-way roads were found to be statistically significant. Other parameters include pedestrians aged over 65 years and under 15 years, drivers under 25 years, male drivers and pedestrians, darkness, heavy vehicles, speed limits exceeding 96 km/h (60 mph), major roads (class A roads), and single carriageway roads are significant in both sections. The study proposes various measures to mitigate the severity of pedestrian-vehicle crashes, such as improving lighting conditions, enhancing pedestrian infrastructure, reducing speed limits in crash-prone areas, and promoting education and awareness among pedestrians and drivers.[Ali23]

2.6 Proactive Big Data Analysis for Traffic Accident Prediction[Fin+20]

In this paper, the authors presented a system for proactive monitoring and forecasting of the risks of road traffic accidents, depending on the influence of external features. To solve this problem, a method for analysis and predictive modeling of changes in the road transport infrastructure has been developed to predict the risk of occurrence and development of destructive events under the influence of external factors. The purpose is to determine, assess, and predict the dynamics of changes in factors that affect the likelihood of the occurrence of risks of accidents, depending on the current situation on the monitored road sections. For predictive risk analysis, information on the parameters of negative events and possible influencing factors obtained from various sources is presented in the form of a spectrum of time series. Comparative analysis of time series of event parameters and factors allows us to identify the causes of incidents and the correlation between factors and events. As factors of influence, meteorological conditions, parameters of automobile and pedestrian traffic on road sections, the state of the road surface, characteristics of road sections, etc. are investigated. The monitoring system is implemented using a multi-agent approach, which involves the use of software agents on photoradar complexes for photo and video registration of road events and mobile communications. Agents solve several tasks of collecting, parsing, consolidating, analyzing, and visualizing big sensory data.[Fin+20]

2.7 Comparison Analysis of Tree Based and Ensembled Regression Algorithms for Traffic Accident Severity Prediction[Muh20]

In this paper, the authors utilized various machine-learning models for accident prediction. Specifically, tree-based ensemble models(Random Forest, AdaBoost, Extra Tree, and Gradient Boosting) and an ensemble of two statistical models (Logistic Regression and Stochastic Gradient Descent) are being used as voting classifiers and are being compared for the prediction of road accident severity. Significant features that are strongly correlated with the accident severity are identified by Random Forest. Analysis proved Random Forest as the best-performing model with the highest classification results with 0.974 accuracies, 0.954 precision, 0.930 recall, and 0.942 F-score using the 20 most significant features as compared to other techniques classification of road accident severity.[Muh20]

2.8 Towards Deep Learning based Traffic Accident Analysis[NNA20]

In this study, the authors propose a framework to store traffic accident data in a big data platform(Cassandra) which will enable fast and accurate analyses and predictions with the help of a deep learning technique to build prediction and classification models from the road accident data. In particular, the models focus on the prediction and identification of hotspots on the roads. [NNA20]

2.9 Data-Driven Urban TrafficAccidentAnalysis and Prediction Using Logit and Machine Learning-Based Pattern Recognition Models[Vah21]

The purpose of this study is to identify the variables affecting the severity of the injury, fatal, and property damage only (PDO) accidents. In this regard, the multiple logistic regression and the pattern recognition type of artificial neural network (ANN) as a machine learning solution are used to recognize the most influential variables on the severity of accidents and the superior approach for accident prediction. Results show that the multiple logistic regression in the forward stepwise method has an R2 of 0.854 and an accuracy prediction power of 89.17%. It turns out that the accidents occurred between 18 and 24 and the KIA Pride vehicle has the highest effect on increasing the severity of accidents, respectively. The most important result of the logit model accentuates the role of environmental variables, including poor lighting conditions alongside unfavorable weather, and the dominant role of unsafe and poor quality vehicles in increasing the severity of accidents. In addition, the machine learning model performs significantly better and has higher prediction accuracy (98.9%) than the logit model. In addition, the ANN model's greater power to predict and estimate future accidents is confirmed through performance and sensitivity analysis.[Vah21]

2.10 An Alternative Method for Traffic Accident Severity Prediction: Using Deep Forests Algorithm[Jin20]

This study employs the UK road safety dataset to propose a novel method for predicting the severity of traffic accidents based on the Deep Forests algorithm. To verify the superiority of their proposed method, several other machine learning algorithm-based perdition models were implemented to predict traffic accident severity with the same dataset, and the prediction results show that the Deep Forests algorithm presents good stability, fewer hyper-parameters, and the highest accuracy under different level of training data volume. To understand the relationship between the features and the target variable, the authors calculated the correlation relationship between all the features and found that most of the characteristic variables are linearly independent, except for weather conditions, road surface, and light condition, the light condition and hour of day, vehicle type, and engine capacity. But, the target feature, accident severity is independent of any of the other features. To deal with this problem, this paper considers a smarter ML approach. In this paper, the Random Forests (RFs) method is adopted to carry out feature selection according to the importance index of each feature. They have selected 8 features selected by the feature selection method and then the Deep Forests Algorithm is adopted to predict the severity of traffic accidents. The ROC had reached 0.93 with an accuracy of 90%.[Jin20]

2.11 Accident Risk prediction based on Machine Learning[B N20]

The main objective of this paper is to decrease the number of road accidents caused due to weather conditions. This system analyzes the amount of risk in terms of the percentage present in that particular area based on the present weather conditions and previously collected datasets. They have used features like latitude and longitude, day of the week, number of days, road class, speed limit, pressure, weather, visibility, wind speed, etc. Also, they have implemented the Logistic Regression and Naive Bayes algorithm to determine the amount of risk in terms of the percentage present in a particular area.[B N20]

2.12 Hybrid feature selection-based machine learning Classification system for the prediction of injury severity in single and multiple-vehicle accidents[Shu22]

The paper presents a hybrid feature selection-based machine learning classification approach for detecting significant attributes and predicting injury severity in single and multiple-vehicle accidents. The authors employed a Random Forests (RF) classifier in conjunction with an intrinsic wrapper-based feature selection approach called the Boruta Algorithm (BA) to find the relevant important attributes that determine injury severity. The influential attributes were then fed into a set of four classifiers to accurately predict injury severity (Naive Bayes (NB), KNearest Neighbor (K-NN), Binary Logistic Regression (BLR), and Extreme Gradient Boosting (XGBoost)). According to BA's experimental investigation, vehicle type was the most influential factor, followed by the month of the year, the driver's age, and the alignment of the road segment. The driver's gender, the presence of a median, and the presence of a shoulder were all found to be unimportant. According to classifier performance measures, XGBoost surpasses the other classifiers in terms of prediction performance. Using the specified attributes, the accuracy of the XGBoost was 82.10% for single-vehicle accidents and 79.52% for multiple-vehicle accidents, respectively.[Shu22]

References

[Hoa17] Fang Chen Hoang Nguyen Chen Cai. "Literature Review on Crash Reporting - Incidents Data". In: (2017).

[B N20] et al. B Nanditha. "Accident Risk prediction based on Machine Learning". In: (2020).

[Fin+20] Anton Finogeev et al. "Proactive Big Data Analysis for Traffic Accident Prediction". In: (2020).

[Jin20] et al. Jing Gan. "An Alternative Method for Traffic Accident Severity Prediction: Using Deep Forests Algorithm". In: (2020).

[Muh20] et al. Muhammad Umer. "Comparison Analysis of Tree Based and Ensembled Regression Algorithms for Traffic Accident Severity Prediction". In: (2020).

[NNA20] Atif Naseer, Mohamed K. Nour, and Basem Y. Alkazemi. "Towards Deep Learning based Traffic Accident Analysis". In: (2020).

[Kom21] et al. Komol MMR. "Crash severity analysis of vulnerable road users using machine learning". In: (2021).

[Vah21] et al. Vahid Najafi Moghaddam Gilani. "Data-Driven Urban TrafficAccidentAnalysis and Prediction Using Logit and Machine Learning-Based Pattern Recognition Models". In: (2021).

[Ras22] Wikasitha Herath Rasika Vijithasena. "Data Visualization and Machine Learning Approach for Analyzing Severity of Road Accidents". In: (2022).

[Shu22] et al. Shuguang Zhang. "Hybrid feature selection-based machine learning Classification system for the prediction of injury severity in single and multiple-vehicle accidents". In: (2022).

[Ali23] et al. Alireza Salehian. "Comparative analysis of pedestrian crash severity at United Kingdom rural road intersections and Non-Intersections using latent class clustering and ordered probit model". In: (2023).

[Aas24] Biroshan Kumar Mahato Aashish Bhandari. "Comprehensive Analysis of Road Traffic Accidents in the United States: Unveiling Trends, Patterns, Predictions, and Road Safety Insights through Big Data Analysis". In: (2024).