# Introduction

Road traffic accidents is a major concern of society due to the loss and injury of millions of lives and billions of dollars of property damage. The big data environment provides opportunity to excavate the factors that would influence the severity of car accidents. In recent years, researchers have studied how various factors affect the car accidents severity, but only few studies has combined all the information, including driver’s information, geographic information, road condition, and crash related information, to analyze factor influence on car accidents severity. Taking advantage of all information might increase the prediction accuracy and help to better understanding the influence significance in a big picture. Moreover, more advanced machine learning models have come out recently, but few studies examine the effect of those models on predicting car accident severity. It would be of special interest to find a better model with higher prediction accuracy and study the important factors that influence the car accident severity with latest dataset. The aim of the experiment is to give instruction for which machine learning model perform better on predicting car accident severity and look for the key factors that would significantly influence the car accident severity.

# Literature Review

In the past few decades, some researchers have applied various models to explore the relationship between accident severity and a raft of explanatory variable of interests. Much of the accident severity research has focused on identifying and evaluating the key factors which could exacerbate an accident and even lead to death. Venkataraman Shankar and Fred Mannering (1996) performed logit analysis of motorcycle accident severity based on a 5-year statewide dataset of Washington, which is different from our focus of car accidents. By leveraging conditional model, they were able to estimate accident severity level based on environmental factors, vehicle characteristics and rider attributes—only for motorcycle, as well as studies by Matthieu (2006) and Mohammad (2014). To date, several studies have investigated how to predict a car accident severity and what is the most lethal factor. In a traffic accident analysis, Miao *et al*. (2005) found that the classification accuracy of different injury level in car accident varied with the modeling approaches. For example, the no injury and the possible injury classes could be best modeled directly by decision tree other than support vector machines or neural network. Overall, their models performed better for fatal and non-fatal classification compared with other classes. Like our study, Miao’s study collected data from the National Automotive Sampling System General Estimates System (NASS GES). However, their data is relatively old and only includes 11 variables in their model. More factors other than the 11 variables may also explain car accident severity, such as weather, gross weight of car and number of occupants, which are not included in Miao’s study. Though they applied some advanced models such as neural network, they did not give a clearly ranking showing which variable is most influential in estimating accident severity. Two years later, Rifaat *et al*. (2007) carried out an analysis to explore how the interaction of multiple factors, such as driver characteristics, roadway features, vehicle types, pedestrian characteristics, etc., will affect accident severity employing ordered probit model. They investigated two-vehicle crashes, single vehicle crashes, and pedestrian accident based on the data collected in Singapore from 1992 to 2001. It demonstrates that vehicle type, road type, collision type, etc., were found to be significantly associated with injury severity. In 2012, Athanaslos *et al*. developed two models for inside and outside urban areas separately. They found that the importance level of features affected accident severity are difference in each area. By focusing on the urban area, Ghulam and co-workers (2012) trained multinomial logistic regression model based on collisions data from urban US highways in Arkansas. Our study filled the gap by adding the category variable of urban/rural area at a country wide level with more advanced algorithms. In 2010, Beshah *et al*. leveraged machine learning to predict the accident severity based on dataset from Ethiopia. Among the models, K-Nearest Neighbors obtained the highest accuracy. Machine learning did a good job on predicting the severity level. The research is helpful for Ethiopian Traffic Agency to improve the road safety. Six years later (2016), Sharaf et al. used an artificial neural network to do prediction of injury severity according to the traffic records (from 2008 to 2013) occurred in Abu Dhabi, and the accuracy for testing data is 74.6%. In 2017, George *et al*. collected car accidents data in Greece to accomplish an investigation of road accident severity per vehicle type. They applied lognormal regression to examine the severity based on three expressions of mathematical functions, which drew the conclusion that crash type, whether and light condition play significant role in an accident. In the study conducted by Sunanda *et al.,* some factors like travel speed and restraint device usage influence the severity, but their interest group is only older drivers (2002). Currently, one study by Kyi *et al*. (2019) applied Functional Resonance Analysis Method (FRAM) and Naïve Bayes algorithm to classify the accident severity based on dataset of road accident in Yangon. Other studies (So, 2010; Mahdi, 2012; Joaquín, 2013; Dimitris, 2018; Luara, 2020; Deyu, 2019) though investigate factors affecting accident severity as well, their investigated area is not US national wide level like our study. Moreover, those studies do not include as many features as ours. Some studies explained accident severity in terms of different traffic condition of rural/urban areas, but, again, their data is collected the road condition of other countries (Ahmad, 2005; Juan, 2011; Natalia, 2019) or partial area of US (Samantha, 2014).

However, those results and models of the studies mentioned above probably are not applicable in United States. They were carried out from completely different dataset from different countries, so the outcomes of machine learning can be highly varied. Even the same area with dataset of different time can cause appreciable variation in models and results. For predicting the accident severity in U.S. area, Delen and co-workers (2017) conducted research and identified the importance of multiple factors by investigating the dataset derived from National Automotive Sampling System General Estimates System (NASS GES). The dataset includes U.S. accident data from 2011 and 2012. By combining data from accident database, vehicle database and person database to obtain a dataset with 29 variables which are relevant and potentially influential to accident severity. There is also a study (Hisaaki *et al*. 2003) utilized the NASS database, but it focused on the FARS dataset which only includes accident with fatality. For the purpose of predicting accident severity not only investigating the fatality events, in our study, we leveraged the database from NASS as well, but we focused on the latest dataset (collected in 2018) and included 226 variables which enable us detecting more potential influencer compared with Delen’s dataset. Different from Delen’s study, we use K-Nearest Neighbors, Random Forest, XGboost and LightGBM at both vehicle-level and person-level in this paper. In our experiments, Random Forest and XGBoost outperformed than other models for prediction the accident severity. We also give a new ranking of importance of factors, which could help the relevant department to improve the road safety in the future.

# Methodology

# Data Description

The National Highway Traffic Safety Administration (NHTSA) is to reduce human toll and damage. NHTSA uses data from many sources, including the Crash Report Sampling System (CRSS). CRSS is a sample of police-reported crashes involving all types of motor vehicles, pedestrians, and cyclists, ranging from property-damage-only crashes to those that result in fatalities. CRSS accident dataset covered accidents including 2018. The dataset has been anonymized by U.S. Department of Transportation, National Highway Traffic Safety Administration and contains 226 variables, capturing personal and the in-transport motor vehicles and the drivers of in-transport motor vehicle information. Note that variables/records within the CRSS dataset have a large number of missing data. Thus, a prerequisite to analyzing the CRSS dataset is a detailed data cleaning/preparation procedure that will facilitate reproducible analyses.

### Data cleaning and preprocessing

Our procedure for cleaning the CRSS dataset was comprised of several steps: The first step focused on changing the values which stand for of 'Not\_reported', 'Unknown, 'None', ‘Not\_Applicable’ to NA by checking these values for each feature. Note that variables/records within the CRSS dataset have a large number of missing values. In our case, we dropped the variables contained too many missing values based on threshold equal to 0.8.

The second step involved determining whether a variable should be encoded. For some categorical variables, we encoded these columns into the numerical order. e.g.: if the original column contains categories 1, 5, 9, 10, we encoded it as 0, 1, 2, 3 correspondingly.

# Results

# Conclusion

# Reference

Kato, H., &amp; Ogata, K. (2003). An Approach for Compatibility Improvement Based on US Traffic Accident Data. SAE Technical Paper Series. doi:10.4271/2003-01-0906

Theofilatos, A., Graham, D., &amp; Yannis, G. (2012). Factors Affecting Accident Severity Inside and Outside Urban Areas in Greece. Traffic Injury Prevention, 13(5), 458-467. doi:10.1080/15389588.2012.661110

Bham, G. H., Javvadi, B. S., &amp; Manepalli, U. R. (2012). Multinomial Logistic Regression Model for Single-Vehicle and Multivehicle Collisions on Urban U.S. Highways in Arkansas. Journal of Transportation Engineering, 138(6), 786-797. doi:10.1061/(asce)te.1943-5436.0000370

Alkheder, S., Taamneh, M., &amp; Taamneh, S. (2016). Severity Prediction of Traffic Accident Using an Artificial Neural Network. Journal of Forecasting, 36(1), 100-108. doi:10.1002/for.2425

Hlaing, K. P., Aung, N. T., Hlaing, S. Z., &amp; Ochimizu, K. (2019). Analysis of accident severity factor in Road Accident of Yangon using FRAM and Classification Technique. 2019 International Conference on Advanced Information Technologies (ICAIT). doi:10.1109/aitc.2019.8921119

So Young Sohn & Hyungwon Shin (2001) Pattern recognition for road traffic accident severity in Korea, Ergonomics, 44(1), 107-117, doi: [10.1080/00140130120928](https://doi.org/10.1080/00140130120928)

Mohammadzadeh, M., Paravar, M., Mirzadeh, A., Mohammadzadeh, J., Mahdian, S. (2015). Seat Belt Usage in Injured Car Occupants: Injury Patterns, Severity and Outcome After Two Main Car Accident Mechanisms in Kashan, Iran, 2012. *Archives of Trauma Research,*, 4(1), -. doi: 10.5812/atr.22203

Abellán, J., López, G., &amp; Oña, J. D. (2013). Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications, 40(15), 6047-6054. doi:10.1016/j.eswa.2013.05.027

Potoglou, D., Carlucci, F., Cirà, A., &amp; Restaino, M. (2018). Factors associated with urban non-fatal road-accident severity. International Journal of Injury Control and Safety Promotion, 25(3), 303-310. doi:10.1080/17457300.2018.1431945

Eboli, L., Forciniti, C., & Mazzulla, G. (2020). Factors influencing accident severity: An analysis by road accident type. Transportation Research Procedia, 47, 449-456. doi:10.1016/j.trpro.2020.03.120

Dissanayake, S., &amp; Lu, J. J. (2002). Factors influential in making an injury severity difference to older drivers involved in fixed object–passenger car crashes. Accident Analysis &amp; Prevention, 34(5), 609-618. doi:10.1016/s0001-4575(01)00060-4

Khorashadi, A., Niemeier, D., Shankar, V., &amp; Mannering, F. (2005). Differences in rural and urban driver-injury severities in accidents involving large-trucks: An exploratory analysis. Accident Analysis &amp; Prevention, 37(5), 910-921. doi:10.1016/j.aap.2005.04.009

Oña, J. D., Mujalli, R. O., &amp; Calvo, F. J. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accident Analysis &amp; Prevention, 43(1), 402-411. doi:10.1016/j.aap.2010.09.010

Islam, S., Jones, S. L., &amp; Dye, D. (2014). Comprehensive analysis of single- and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. Accident Analysis &amp; Prevention, 67, 148-158. doi:10.1016/j.aap.2014.02.014

Casado-Sanz, N., Guirao, B., & Gálvez-Pérez, D. (2019). Population ageing and rural road accidents: Analysis of accident severity in traffic crashes with older pedestrians on Spanish crosstown roads. Research in Transportation Business & Management, 30, 100377. doi:10.1016/j.rtbm.2019.100377

Lapparent, M. D. (2006). Empirical Bayesian analysis of accident severity for motorcyclists in large French urban areas. Accident Analysis & Prevention, 38(2), 260-268. doi:10.1016/j.aap.2005.09.001

Shaheed, M. S., & Gkritza, K. (2014). A latent class analysis of single-vehicle motorcycle crash severity outcomes. Analytic Methods in Accident Research, 2, 30-38. doi:10.1016/j.amar.2014.03.002

Beshah, T., & Hill, S. (2010). Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. *AAAI Spring Symposium: Artificial Intelligence for Development*.

George, Y., Athanasios, T., & George, P. (2017). Investigation of road accident severity per vehicle type. Transportation Research Procedia, 25, 2076-2083. doi:10.1016/j.trpro.2017.05.401

Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. Journal of Transport & Health, 4, 118-131. doi:10.1016/j.jth.2017.01.009

Chong, M.M., Abraham, A., & Paprzycki, M. (2005). Traffic Accident Analysis Using Machine Learning Paradigms. *Informatica (Slovenia), 29*, 89-98.

Rifaat, S. M., & Chin, H. C. (2007). Accident severity analysis using ordered probit model. Journal of Advanced Transportation, 41(1), 91-114. doi:10.1002/atr.5670410107

Shankar, V., & Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. Journal of Safety Research, 27(3), 183-194. doi:10.1016/0022-4375(96)00010-2