A Data-driven Framework to Identify Human-Critical Autonomous Vehicle Testing and Deployment Zones

HM Abdul Aziz, Ph.D.
(Corresponding Author)

Department of Civil Engineering
Kansas State University
Manhattan, Kansas, United States
azizhusain@ksu.edu

A M Hasibul Islam

Graduate Researcher
Department of Civil Engineering
Kansas State University
Manhattan, Kansas, United States
afzalhasib91@ksu.edu

ABSTRACT

We proposed a data-driven framework that leverages machine learning and econometric modeling techniques to investigate autonomous vehicle (AV) involved crashes and maps human-critical safety factors to operational design domains. The goal is to assist the infrastructure owner-operators in identifying human-critical AV-testing-and-deployment zones based on data-driven insights from both AV-testing data (e.g., California Department of Motor Vehicle AV crash reports) and historical crash data involving only human drivers. First, we analyzed AV crash data collected from the CA DMV website for May 2018 to December 2020 using ML-based and econometric models incorporating attributes such as weather, lighting condition, road surface condition, vehicle miles traveled, and collision type. Later we use the findings to demonstrate the framework's applicability for New York City crash data at the Zip Code level (2012 -2021).

CCS CONCEPTS

• Computing Methodologies • Applied Computing

KEYWORDS

Automated Vehicles, Operational Design Domain, Traffic Safety, Human Factors

1 Introduction

Reliable and trustworthy deployment of autonomous vehicles (AVs) warrants rigorous safety assessment by the Infrastructure Owner-Operators (IOOs). The goal is to build public confidence in the safety and efficiency of the Automated Driving System (ADS) facilitated by the AVs (SAE level 3 and higher). The high-dimensional stochastic traffic environment makes it challenging

(if not infeasible) to generate and validate safety assessment data for all possible scenarios [8]. According to National Highway Traffic Safety Administration (NHTSA)[15], the ADS testing and deployment entities should demonstrate (to the public—States and consumers) four aspects in their respective Voluntary Safety Self-Assessments (VSSA): (a) safety considerations of ADSs, (b) communicating and collaborating with the State Department of Transportation (DOT) units, (c) support to self-establish of industry safety norms, and (d) development of public trust, and acceptance through transparent testing and deployment of ADSs. Operational Design Domain (ODD) plays a significant role in assessing the safety of the AVs. The SAE J3016 defines ODD as *Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of specific traffic or roadway characteristics*[22]. The ODD is generally defined by the AV manufacture (or the ADS system developer) that describes (at the minimal) technical capabilities of an ADS (algorithms, sensing, use cases, and business models), operational boundaries, and the risk management strategies for the deployed ADS [21]. An ODD is also constrained by the regulatory policies and restrictions within a geographic region imposed by the IOOs as a function of road classifications (accessibility and mobility), safety regulation for vulnerable road users, and pre-established knowledge regarding crash contributing factors.

While AV testing is the key to gaining confidence in safety assurance, IOOs are constrained by budgetary and legal policies for SAE Level 3+ deployments. Entire twenty-nine states have established guidelines and legislation regarding AV operations (including testing) on roads [9]. Several states, including California, Michigan, and Florida, have ongoing AV testing and deployment projects [30,31] following NHTSA's testing guidelines (led by ADS vendors such as *Waymo*, *GM Cruise*, *Zoox*).

The AV industry and IOOs have acknowledged that it will be challenging to define and regulate ODDs for ADS features where AVs and human-driven vehicles will share the same road facilities. Now, it is possible to leverage extensive historical crash data (human-only) from the existing road networks to understand the role human-errors in certain types of traffic crashes. Particularly,

This work is licensed under a Creative Commons Attribution-NonDerivs International 4.0 License.

IWCTS'21, November 1, 2021, Beijing, China

© 2021 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-9117-7/21/11.

<https://doi.org/10.1145/3486629.3490692>

the focus should be on the types of crashes those occur in both networks—human-only (existing) and mixed-autonomy (future roads carrying both human drivers and AVs). Waymo recently analyzed historical fatal crash data (2008 through 2017) for one of its current ODDs and leveraged counterfactual simulations to demonstrate the safety benefits of ADS [24]. This analysis underscored the role of *imperfect* human drivers dictating the bounds and uncertainty of safety benefits of ADSs. For instance, some fatal crashes (e.g., rear-end crashes with a human driver at fault) could not be prevented even if we have an SAE level 4+ AV responding with the optimal risk-mitigation strategies (Dynamic Driving Task fall back and stabilization). Thus, it is important for the IOOs to carefully identify the deployment zones where the human drivers are more likely to contribute to unavoidable crashes involving an AV—***the human-critical AV deployment zones***. We define a human-critical AV testing and deployment zone as the location (or analysis area) with a high probability of AV-human crash events with human driver(s) as the initiator—primarily responsible for the event.

To identify the potential human-critical AV deployment zones (mixed environment of AVs and human drivers), we propose a data-driven framework that leverages AV-testing and the historical crash data specific to geographic ODDs. The goal is to assist the IOOs in planning for AV testing and deployment, accounting for human-driving imperfections and identifying testing-and-deployment geofences with minimal risks concerning human factors in a mixed driving environment.

Figure 1: Data-driven Framework to Identify/Prioritize AV Deployment and Testing Zones

Figure 1 describes the proposed data-driven framework. Two primary data sources required are the AV-testing database (including all types of events—crashes, near-misses, disengagement, and DDT-fallback actions) and the historical traffic crash data for the region of interest with higher spatial resolution such as census tracts, intersections, and roadway segments.

Consider the case where rear-end crashes are found to be highly probable in the AV-testing data and correlated with human-driving errors (as the initiator) in a mixed environment. To minimize the risk associated with these incidents, an IOO needs to identify the roadway segments or signalized intersections where rear-end crashes are highly probable. Given that we will not have sufficient testing data from mixed human-AV networks, one approach is to analyze the existing human-only crash data focusing on the features found in AV-testing data. Using crash analysis techniques (either econometric or machine-learning based), we can build a spatial-temporal mapping between the features present in the AV-related and historical human-caused crashes or similar events.

To this end, our goal is to develop a data-driven framework that will build a spatial-temporal mapping between the crash (or similar events) contributing factors found in the human-only historical crash data and the correlated features observed in AV-testing data. This will allow the IOOs to identify the ***human-critical AV deployment zones*** accounting for spatial-temporal heterogeneity in the human-contributing factors.

We demonstrate the applicability of the proposed framework using the collection of crash reports of the California Department of Motor Vehicles (CA DMV) from May 2018 to December 2020 [5]. Using the findings from this dataset, we identify the human-critical AV deployment zones in New York City, NY. However, we acknowledge that this transferability is weak and will serve as an initial direction for the IOOs when sufficient AV-testing data are not available.

2 Related Works: Crash data Analysis Methods

The literature on human-only crash-data studies exploring and identifying the contributing factors is rich and well documented [13, 23]. It is possible to build geographically segmented models for crash frequency and severity (e.g., at borough level in New York City, NY [3]; at signalized intersections [19]) using state-of-art modeling techniques such as random parameter logit models [1] and Bayesian models [29]. Besides the econometric approaches, data mining and machine learning models have gained popularity in recent years for analyzing crash data. Studies attempted to identify the contributing factors, the relative importance of the features, and prediction power [4]. Huang et al. [10] used classification and decision trees to evaluate the effect of road terrain, environmental conditions, driver characteristics on the crash injury outcome. Jeong et al. [11] used multiple classification trees (e.g., gradient boosting, decision tree, neural network) to classify the injury rate with higher success. One challenge with the machine learning (ML) based prediction models is their lack of inference capability. Mannering et al. [23] stated that ML models have a high predictive capability but very low inference capability regarding crash severity modeling and

prediction. In comparison, causal inference-based models have high inference capability but deficient predictive capability.

AV-testing data are scarce, and most studies are based on the CA DMV-provided crash data [5]). Wang et al. [26] used logistic regression and the Classification and Regression Tree (CART) model using California DMV's 2014 – 2018 AV crash data. They found that the crash severity increases when the AV unit is primarily responsible for the event. Subhashish et al. [6] used a Bayesian latent class approach for the CA DMV data (2014 to 2019) to identify six classes of collision patterns. Favaro et al. [7] reported speed distribution for vehicles involved in AV crashes and visual reconstruction of AV-related crashes. Xu et al. [28] found a correlation between the severity levels of AV crash with collision location, rear-end collision, one-way road, and driving mode (auto vs. conventional).

The studies mentioned above explored and investigated limited AV testing data to find the contributing factors to the occurrence and severity of an AV-crash-related event. However, none of these studies looked at the possible mapping of features from AV-crash data to historical human-drivers crash factors to the best of our knowledge. Our contribution here is the proposed data-driven framework for spatial-temporal mapping of crash-correlation factors that will assist the IOOs in prioritizing AV testing and deployment zones and developing regulations and restrictions for the ODDs in the region of interest. We also integrated ML and econometric techniques into the modeling process. The ML techniques are used to explore and identify the contributing factors (use feature importance), and the variables with high importance are used in the model specification of the econometric models. Finally, we utilized the findings from the econometric model to identify human-critical AV deployment zones for New York City, NY.

3 AV Crash Data Description

Crash reports of AVs were extracted from the website of the California DMV [5]. Individual crash reports starting from May 2018 to December 2020 were combined into a single data file for analysis with a total of 189 AV-related crash data points. The crash reports before May 2018 were not considered for this research, as those reports lacked the records on some of the newly added variables. Crash events were categorized into two outcome groups: injury and no injury—used as the dependent variable in the model specification. Among the total 189 crash incidents, 29 led to injuries, and 160 produced no injury crashes. Figures 2 and 3 show the injury outcome distribution based on driving mode and type of collision.

To understand the correlation between vehicle-miles-traveled (VMT) of the AVs and the occurrence of a crash (or incident), we categorized the crash-events into four groups: *Category A* (VMT is smaller than 100,00 miles), *Category B* (VMT is greater than 100,000 miles but smaller than 500,000 miles), *Category C* (VMT is greater than 500,000 miles but smaller than 1000,000 miles), and *Category D* (VMT is greater than 100,000 miles). Figure 4 shows

the frequency distribution of different categories of annual vehicle miles traveled with injury outcomes. We use the average VMTs reported by the manufacturers and mapped the values to the year when the crash event occurred. For instance, if a Waymo AV experienced a crash in 2019, the VMT represents the total testing VMT till mid-2019. It is found that *Category B* and *Category C* VMTs have produced all the injury resulting crashes; *Category C* produced higher counts compared to *Category B*.

The AV-testing dataset contains information including weather conditions (e.g., rain, cloud, snow, sunny), visibility conditions (e.g., dusk, dawn, daylight), movement preceding to collision (e.g., stopped at traffic, turning right, turning left), type of collision (e.g., rear-end, sideswipe), and road surface conditions (e.g., dry, wet). Other significant variables found from descriptive statistics are clear weather, sideswipe type of collision, dark with a streetlight, cloudy weather, no unusual activity on the roadway, and AV stopped in traffic before the collision. Figure 5 shows the injury outcome distribution by weather conditions. The descriptive statistics are similar to the findings of Xu et al. [9], who analyzed CA DMV data for January 2015 to May 2018.

Figure 2: Distribution of AV crash outcomes by driving mode

Figure 3: Distribution of AV crash outcomes by collision types

Figure 4: Distribution of AV crash outcomes by VMT categories

Figure 5: Distribution of AV crash outcomes by weather conditions

4 Data Analysis Methodologies

This section briefly describes the machine-learning (ML) and econometric methods used to analyze the CA-DMV AV crash data in this research.

4.1 Feature Selection Using ML Models

To select the best features from the dataset, a Python-based in-built class *Feature Importance* was used (Scikit-learn library (ensemble)). Feature importance is a method of listing features in terms of a given score for the dataset's features. The higher the score, the more the influence of that feature on the outcome. We have used three ML algorithms that use the results from the feature selection algorithm.

4.1.1 K Nearest Neighbor (KNN):

KNN is one of the fundamental classification methods which classifies new data points based on similarity measures (distance functions). This classifier is based on the Euclidean distance between a test sample and a specified train sample. Let x_i be an input sample with s features ($x_{i1}, x_{i2}, \dots, x_{is}$), n be the total number of input samples ($i = 1, 2, \dots, n$), and s is the total number of features ($j = 1, 2, \dots, s$). Then the Euclidean distance between x_i and x_t is [18]:

$$D(x_i, x_t) = \sqrt{(x_{i1} - x_{t1})^2 + (x_{i2} - x_{t2})^2 + \dots + (x_{is} - x_{ts})^2} \quad (1)$$

4.1.3 Decision Tree:

A decision tree [20] algorithm consists of nodes and branches. Each internal node represents a "test" on a specified attribute, each branch represents the outcome of the test, and each leaf node represents a class label. To construct a decision tree, a set X of training data is taken, which has C classes ($C_1, C_2, \dots, C_C$) [25]. X will be a leaf if X has cases of just one single class. X will be a leaf as well if it has no cases. The class adjacent to this leaf will become associated with the major class of its parent node. For more than one class, a test is performed on the training data based on some specified feature attributes x_i . From this, n number of subsets will be created by splitting X where n is the number of outcomes over x_i .

4.1.4 Random Forest:

Random forest is a widely used supervised learning algorithm that applies the ensemble learning method for regression. It operates by constructing numerous decision trees and outputting the mean of the classes as the prediction of all the trees. The random forest is a collection of tree predictors $p(x; \varphi_k)$, $k = 1, 2, \dots, k$ where x is the observed input vector of length l with adjacent random vector X and φ_k represent random vectors which are independent and identically distributed (*i.i.d*) [14]. It is assumed in this case that the observed training data is independently drawn from the combined distribution of (X, Y) . For regression, the random forest prediction is expressed as below:

$$\bar{p}(x) = \left(\frac{1}{K}\right) \sum_{k=1}^K p(x; \varphi_k) \quad (2)$$

The above-mentioned algorithms are applied through the Python Scikit-learn library. The dataset was divided into train and test set by using the 75:25 ratio. A random state of 10 was used in all the models.

4.2 Crash Severity: Logit Models

We follow the (random-utility-based) multinomial logit model specifications for crash severity [2,27]. The model estimates the probability of severity outcome: injury and no-injury. A mixed-logit (random-parameter) model was also estimated; however, no variable was found to be random (not justified); therefore, not reported in this research.

5 Results: Crash Contributing Factors

5.1 Feature Importance

Using the feature importance metrics, 25 variables were selected. It is found that *Category C VMT* received the highest score among the features—an importance score of 0.143477. Figures 6 - 8 provide a graphical representation of these variables with respective feature importance scores. *Driving mode (auto)* and *AV proceeding straight* are the next two most important variables.

Figure 6: Feature Importance Scores (Rank 1-8)

Figure 7: Feature Importance Scores (9-16)

Figure 8: Feature Importance Scores (Rank 17-25)

5.2 Results from ML Models

Next, these selected features were tested with three models: KNN, decision trees, and random forest classifiers. Data points were divided into train sets and test sets by the 75:25 ratio. The results are shown in Table 1 and Figure 9.

Table 1: Results from ML Models for Selected 25 Features

Statistics		Decision Tree	Random Forest	KNN
Accuracy		0.86	0.92	0.90
Precision	Injury	0.36	0.5	0.4
	No injury	1	0.98	0.95
Recall	Injury	1	0.75	0.5
	No injury	0.84	0.93	0.93
F1 score	Injury	0.53	0.6	0.444
	No injury	0.91	0.95	0.94
Support	Injury	4	4	4
	No injury	44	44	44
Confusion matrix		[[37 7] [0 4]]	[[41 3] [1 3]]	[[41 3] [2 2]]

Figure 9: Model comparison for selected best features

From Table 1, it is found out that the model prepared with a random forest classifier produces the best accuracy (92%) with the highest precision (0.5) and F1 score for both injury and no injury outcomes. From the confusion matrices, it is found out that the random forest model was able to correctly predict 41 out of 44 no injury crashes and 3 out of 4 injury crashes. Only the decision tree model predicted all the injury resulting outcomes correctly despite its poor accuracy in predicting the no injury crashes (37 out of 44). Overall, the random forest model showed better prediction accuracy.

With caution, it can be concluded that the feature importance method was able to successfully extract the most influential 25 features from the dataset. From the variable list of best features, it is found that VMT had a total of 3 variables out of 4 as influential ones. All the variables in the light condition category were also found as highly influential. Regarding types of collision, *broadside and rear-end type of collisions* were found to be influential, supporting previous findings [7][6]. The driving mode was also found as influential, supporting previous research [26].

5.3 Estimates from Logit Models

Leveraging the feature importance values, we specified the utility functions of the multinomial logit models, and Table 2 reports the estimation results (Loglikelihood value: -66.67 and adjusted Rho-square 0.1547). Results indicate that crash events occurring with daylight conditions decrease the probability of injury-outcome compared to no-injury. The estimate for Rear-end collision indicator is positive, indicating a higher likelihood of a resulting injury when an AV is involved in a rear-end collision. It is interesting to find *Category-C VMT* increases the likelihood of an injury outcome compared to non-injury—this may reflect the uncertainty in technological readiness of the AV units (not enough miles driven). We acknowledge that the dataset is too small to make any strong inferences; however, the estimates will point us to the critical factors associated with AV crashes.

Table 2: Results from the AV-Crash severity (Logit) model

Variable	Class: Injury		Class: No-Injury	
	Estimate	Z-Score	Estimate	Z-Score
Constant			3.73	3.96
Daylight indicator (1 if the incident occurred in daylight conditions, 0 otherwise).	-0.498	-1.11		
Vehicle-Miles-Traveled indicator (1 if the testing miles is greater than 500K but lower than 1M at the time of the incident)	2.204	3.82		
Rear-end collision indicator—Human initiator (1 if the collision was rear-ended, 0 otherwise)	1.168	1.48		
Lane-changing movement indicator (1 if the non-AV was making a lane-changing maneuver during the event, 0 otherwise)			0.913	1.32
AIC	143.3			
Number of observations	189			

6 Mapping of AV-Crash Factors to Human-Critical Deployment Zones

In a *data-perfect* world, we should conduct the crash factor analysis using AV-testing data from networks that are reasonably representative of the geographic area of interest—future deployment area. However, the publicly available AV-testing data are scarce and limited to certain conditions (semi-controlled

environment). We use the rear-end crash criterion to demonstrate our framework (figure 1) to the minimal form to identify human-critical AV-testing and deployment zones. Our ML-based and econometric models results indicate that rear-end collisions are correlated with AV-crashes, resulting in injury outcome—high feature importance score and statistical significance.

We chose New York City, NY, as the geographic area of interest and applied the framework (figure 1) for historical human-only crash data containing about 1.75 M records from 2012 until April 2021 [17]. The dataset has information from all police-reported motor vehicle collisions in New York City, NY. For each crash record, we extracted information about the location (Zip Code) and the primary contributing factors to the collision. Our primary interest was on the identified factor (as documented by the police)—“*Vehicle Following Too Closely*.” For each Zip code, we filtered the crashes with this primary reason because this is an accepted surrogate factor for rear-end collisions [12]. A total of 178 Zip Codes were mapped with rear-end collisions [16].

Figure 8: Percent of Rear-End Crashes at Zip Code Level in NYC (2012 - 2021)

Figure 9: Actual Counts of Rear-End Crashes at Zip Code Level in NYC (2012 - 2021)

Our analysis showed that rear-end AV crashes are potentially critical when caused and initiated by human drivers. Thus, the Zip Codes with more collisions with “following too close” (the surrogate measure) as the primary cause can be labeled as a critical AV-deployment zone. Since human drivers in this geographic area are more likely to initiate rear-end crashes, the probability of having a similar AV crash is higher. Of course, this only provides a qualitative direction rather than a rigorous statistical insight; however, this would be a decision support tool for the IOOs to rank the potential deployment zones. Figure 8 and Figure 9 show the percentage and actual counts of rear-end crashes in NYC at Zip Code levels. This map supported by the AV-crash analysis will allow the IOOs to identify the proper testing and deployment zones accounting for the human-driving attributes—*following too closely*—a contributing factor for the rear-end crashes. The maps (Figures 8 and 9) show that the counts and percentages do not reflect the same level of criticality. The normalized value (percentage of the total types of crashes) underscores the risk associated with the AV-deployment compared to *hot zones* where any traffic crash (in addition to rear-end crashes) is more probable. In other words, a particular geographic region may be labeled as a low-risk traffic crash zone; however, it may be critical for AV deployment because of the presence of specific human factors associated with AV-related crashes.

7 Conclusion

We proposed a data-driven framework for IOOs to identify human-critical AV-testing and deployment zones that provides a mapping between the contributing factors in AV-testing data (crash or similar events) and the historical crash data involving human-only drivers. AV crash data collected from the CA DMV website for the period of MAY 2018 to Dec 2020 were used to develop both ML-based and econometric models using attributes such as weather, lighting condition, road surface condition, vehicle miles traveled, and collision type. One common feature found from both ML and econometric approaches is the high correlation of rear-end collision with AV-crash injury occurrence, which is later used to demonstrate the framework for NYC crash data at Zip code level (2012 -2021).

7.1 Limitations and Future Work

A major challenge with AV crash analysis is the lack of data—our research involves only 189 data points. The low sample size may have created some bias in the results. Nevertheless, the most important features can be categorized using *feature importance* values and incorporated with ML and econometrics modeling techniques. This challenge could be addressed by constructing virtual crash scenarios in augmented reality to generate synthetic data. Also, more data on AV crashes and the proposed framework will provide a significant advantage in advancing the AV deployment accounting for minimal safety risks. The proposed framework can effectively label the *human-critical AV-testing-and-deployment zones* leveraging ML-based and econometric models. As future work, we plan to develop mixed-logit models

accounting for spatial and temporal heterogeneity instead of descriptive statistics used in the current research. Further, the consideration of temporal distribution (peak vs. off-peak), the perception-reaction time distribution of the drivers, the frequency of events correlation with distraction, and the contribution of geometric design will be integrated. This will assist in establishing the regulations and restrictions of a potential ODD before AV testing and deployment.

REFERENCES

- [1] P C Anastasopoulos and F L Mannering. 2011. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. *Accid. Anal. Prev.* 43, 3 (2011), 1140–1147. DOI:https://doi.org/10.1016/j.aap.2010.12.024
- [2] H.M.A. Aziz, S.V. Ukkusuri, and S. Hasan. 2013. Exploring the determinants of pedestrian-vehicle crash severity in New York City. *Accid. Anal. Prev.* 50, (2013). DOI:https://doi.org/10.1016/j.aap.2012.09.034
- [3] H M Abdul Aziz, Satish V Ukkusuri, and Samiul Hasan. 2013. Exploring the determinants of pedestrian-vehicle crash severity in New York City. *Accid. Anal. Prev.* 50, (January 2013), 1298–309. DOI:https://doi.org/10.1016/j.aap.2012.09.034
- [4] CA DMV. 2020. Report of Traffic Collision Involving an Autonomous Vehicle. 316, (2020), 2–4.
- [5] Calofronia DMV. 2021. Autonomous Vehicle Collision Reports. *Department of Motor Vehicles*.
- [6] Subasish Das, Anandi Dutta, and Ioannis Tsapakis. 2020. Automated vehicle collisions in California: Applying Bayesian latent class model. *IATSS Res.* 44, 4 (2020), 300–308. DOI:https://doi.org/10.1016/j.iatssr.2020.03.001
- [7] Francesca M. Favaro, Nazanin Nader, Sky O. Eurich, Michelle Tripp, and Nareesh Varadaraju. 2017. Examining accident reports involving autonomous vehicles in California. *PLoS One* 12, 9 (2017), 1–20. DOI:https://doi.org/10.1371/journal.pone.0184952
- [8] Shuo Feng, Xintao Yan, Haowei Sun, Yiheng Feng, and Henry X. Liu. 2021. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nat. Commun.* 12, 1 (2021), 1–14. DOI:https://doi.org/10.1038/s41467-021-21007-8
- [9] Bryan Gibson. 2017. *Analysis of Autonomous Vehicle Policies*. Retrieved from https://uknowledge.uky.edu/krc_researchreports/1568/
- [10] Helai Huang, Yunying Peng, Jie Wang, Qizhang Luo, and Xiang Li. 2017. Interactive risk analysis on crash injury severity at a mountainous freeway with tunnel groups in China. *Accid. Anal. Prev.* 111, (November 2017), 56–62. DOI:https://doi.org/10.1016/j.aap.2017.11.024
- [11] Heejin Jeong, Youngchan Jang, Patrick Bowman, and Neda Masoud. 2018. Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. *Accid. Anal. Prev.* 120, (August 2018), 250–261. DOI:https://doi.org/10.1016/j.aap.2018.08.025
- [12] Suzanne E Lee, Eddy Llaneras, Sheila G Klauer, and Jeremy Sudweeks. 2007. *Analyses of Rear-End Crashes and Near-Crashes in the 100-Car Naturalistic Driving Study to Support Rear-Signaling Countermeasure Development*.
- [13] Dominique Lord and Fred Mannering. 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transp. Res. Part A Policy Pract.* 44, 5 (2010), 291–305. DOI:https://doi.org/10.1016/j.tra.2010.02.001
- [14] R Mark and Mark R Segal. 2004. *Machine Learning Benchmarks and Random Forest Regression*. (2004).
- [15] NHTSA. 2017. *Automated Driving Systems: A Vision for Safety 2.0*. Washington DC, United States. Retrieved from https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf
- [16] NYC Opendata. 2021. New York City Zip GIS Shapefile.
- [17] NYC OpenData. 2021. NYC Motor Vehicle Crash Data (2012 - 2021).
- [18] Leif E Peterson. 2021. K-nearest neighbor Characteristics of kNN. 4, 2009 (2021), 1–12. DOI:https://doi.org/10.4249/scholarpedia.1883
- [19] Srinivas S Pulugurtha and Venkata R Sambhara. 2011. Pedestrian crash estimation models for signalized intersections. *Accid. Anal. Prev.* 43, 1 (January 2011), 439–46. DOI:https://doi.org/10.1016/j.aap.2010.09.014
- [20] J. R. Quinlan. 1993. *C4.5: programs for machine learning*. San Mateo, Calif.: Morgan Kaufmann Publishers, ©1993.
- [21] SAE Industry Technologies Consortia. 2020. *AVSC Best Practice for Describing Operational Design Domain: Conceptual Framework and Lexicon*.
- [22] SAE J3016. 2021. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*.
- [23] Peter T Savolainen, Fred L Mannering, Dominique Lord, and Mohammed A Quddus. 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accid. Anal. & Prev.* 43, 5 (2011), 1666–1676. DOI:https://doi.org/10.1016/j.aap.2011.03.025
- [24] John M Scanlon, Kristofer D Kusano, Tom Daniel, Christopher Alderson, Alexander Ogle, and Trent Victor. 2021. Waymo Simulated Driving Behavior in Reconstructed Fatal Crashes within an Autonomous Vehicle Operating Domain. *Waymo LLC*. (2021), 1–24.
- [25] Gary Stein, Annie S Wu, and Kien A Hua. 2005. Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection. (2005), 136–141.
- [26] Song Wang and Zhixia Li. 2019. Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches. *PLoS One* 14, 3 (2019), 1–16. DOI:https://doi.org/10.1371/journal.pone.0214550
- [27] Simon P Washington, Matthew G Karlaftis, and Fred L Mannering. 2011. Statistical and econometric methods for transportation data analysis. (2011).
- [28] Chengcheng Xu, Zijian Ding, Chen Wang, and Zhibin Li. 2019. Statistical analysis of the patterns and characteristics of connected and autonomous vehicle involved crashes. *J. Safety Res.* 71, (November 2019). DOI:https://doi.org/10.1016/j.jsr.2019.09.001
- [29] X. Zhan, H.M. Abdul Aziz, and S.V. Ukkusuri. 2015. An efficient parallel sampling technique for Multivariate Poisson-Lognormal model: Analysis with two crash count datasets. *Anal. Methods Accid. Res.* 8, (2015). DOI:https://doi.org/10.1016/j.amar.2015.10.002
- [30] Autonomous Vehicles - California DMV. Retrieved August 23, 2021 from <https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/>
- [31] *Center of Automotive Research. Automated Vehicle Industry Activities in Michigan. MDOT REQ. NO 1259. Michigan Department of Transportation, 2017.*