Accident Analysis and Prevention 174 (2022) 106769

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

What can we learn from autonomous vehicle collision data on crash severity? A cost-sensitive CART approach

Siying Zhu, Qiang Meng *

Department of Civil and Environmental Engineering, National University of Singapore, Singapore 117576, Singapore

ARTICLE INFO

Keywords:
Autonomous vehicle
AV crash severity
Class imbalance
Classification and regression tree

ABSTRACT

Autonomous vehicles (AVs) are emerging in the automobile industry with potential benefits to reduce traffic congestion, improve mobility and accessibility, as well as safety. According to the AV collision data managed by the California Department of Motor Vehicles (DMV), however, the safety issue of AVs has continuously been a concern. This paper aims to learn the contributing factors to AV crash severity from the latest 3-year AV collision data. To achieve the objective, we develop an AV crash severity classification tree with the possible contributing factors by the cost-sensitive classification and regression tree (CART) model, which can deal with the class imbalance issue raised from the AV collision dataset. Our results show that the main factors affecting AV crash severity level include manufacturer, facility type, movement preceding collision, collision type, light condition and year. These findings could provide useful insights for traffic engineers or AV manufacturers to raise effective counter measures or policies to mitigate AV crash severity.

1. Introduction

In recent years, autonomous vehicles (AVs) that utilise artificial intelligence and advance communication techniques to assist human drivers in adjacent situation examination and vehicle speed or steering control operation have become increasingly popular in the automobile industry (Das et al., 2020; Chen et al., 2019). It has been predicted that AVs can bring potential benefits such as reduced vehicular emission (Liu et al., 2019), improved mobility and accessibility (Chen et al., 2020b), increased road capacity (Chen et al., 2022), reduced operation costs (Chen et al., 2021a), reduced road congestion (Chen et al., 2021b), as well as safer road (Song et al., 2021; NSTC, 2020). To ensure AV safety, extensive on-road testing has been conducted on both closed courses and public roads by both private and public organisations (Koopman and Wagner, 2018). Nevertheless, in spite of the development in AV technology, the safety issue of AV becomes a concern of manufacturers, government agencies, and the general public, as AV on-road testing has been carried out (Wang and Li, 2019a). The California Department of Motor Vehicle (DMV) has requested the manufacturers testing AVs on California public roads to report collisions since 2014, providing crucial information about AV crash related factors. According to DMV (2021), more than one out of four AV-involved crashes result in human injuries from 2018 (Oct) to 2021 (Oct), causing economic and social cost. To enhance the AV safety level and mitigate the related cost, it is critical to understand the factors that contribute to more severe AV crashes.

Extensive research has been conducted to investigate the contributing factors to crash severity outcome. Methodologically, regression models such as logistic regression and its variants have been commonly applied for crash severity modelling (Prati et al., 2017; Mujalli and de Ona, 2013); in particular, Xu et al. (2019) and Wang and Li (2019b) have addressed AV crash severity issue. However, most regression models rely on statistical assumptions and pre-defined relationship between independent variables and target variable (Chang and Wang, 2006). In recent years, non-parametric models became increasingly important to search for structures, commonalities, and hidden patterns or rules of datasets (Han et al., 2011). In particular, the classification and regression tree (CART) is a typical non-parametric model that widely used in crash severity analysis to capture the underlying non-additive relationship between crash severity level and crash contributing factors (Huang et al., 2018; Chang and Wang, 2006; Prati et al., 2017; Jung et al., 2016; Chang and Chien, 2013), where satisfiable results and interpretability have been demonstrated.

Moreover, the AV crash dataset usually suffers from a critical issue of class imbalance, as the size of more severe AV crashes is much smaller than that of the less severe ones. As a result, the AV crash severity classification model could classify the less severe AV crashes more accurately, if the AV crash data class imbalance issue is not properly tackled. As the size of AV crash dataset is small, to avoid data distortion and overfitting which could be caused by data resampling (Chawla,

* Corresponding author.
E-mail address: ceemq@nus.edu.sg (Q. Meng).

https://doi.org/10.1016/j.aap.2022.106769
Received 16 August 2021; Received in revised form 17 April 2022; Accepted 2 July 2022
Available online 18 July 2022

2009; Liang and Zhang, 2012), the cost-sensitive learning method is applied in this study to handle the class imbalance issue. The method emphasises more on the minority more severe AV crashes by assigning higher weight or misclassification cost to the class, whereas the distribution of the original AV crash dataset is left unchanged. Therefore, the focus of this study is to investigate the relationship between AV crash severity and influencing factors using the cost-sensitive CART algorithm, based on the AV crash data of California DMV.

1.1. Relevant studies

Many researchers have focused on the analysis of AV safety using AV road test data. For instance, studies have addressed the comparison between AV and conventional human-driving vehicle (HV) crash rate (Goodall, 2021; Favarò et al., 2017; Schoettle and Sivak, 2015; Banerjee et al., 2018); AV disengagement (Wang and Li, 2019a; Boggs et al., 2020a; Dixit et al., 2016; Favarò et al., 2018); descriptive analysis (Favarò et al., 2017), statistical analysis (Das et al., 2020; Boggs et al., 2020b), crash sequence analysis (Song et al., 2021), and text analysis (Alambeigi et al., 2020; Boggs et al., 2020b) of AV collision report; AVs' safety reliability (Li and Zhai, 2019; Kalra and Paddock, 2016). Nevertheless, limited research has been dedicated to AV crash severity analysis. Leilabadi and Schmidt (2019) has analysed AV damage with descriptive statistics and the result showed that the majority of rear end collision were with minor damage to AV. Xu et al. (2019) applied the ordinal logistic regression method for connected and autonomous vehicle (CAV) crash severity modelling, which relies on assumptions such as normal distribution of random error term and proportional odds assumption; classification and regression tree (CART) model was further applied to classify various types of collisions. Wang and Li (2019b) have examined AV crash severity with ordinal regression model, and found that the driving mode, collision location, roadside parking, rear-end collision, and one-way road are the significant influencing factors. Although the regression models mentioned above are able to investigate the effect of multiple factors on AV crash severity, there could be several issues: Firstly, as mentioned above, these models relies on strict statistical assumptions which are difficult to satisfy in practice (Prati et al., 2017). Secondly, the interaction among various variables that occur in a complex form can hardly be catered for (Weng and Meng, 2012; Yan et al., 2010).

Another typical issue for crash severity modelling is to learn from the class-imbalanced dataset, as the majority of crashes belong to the less severe category while the more severe crashes are rare. When balanced data is presumed, the crash severity analysis model would be biased toward the majority class, whereas the minority more severe crashes are often misclassified. Thereby, the class imbalance issue has been a growing concern in the field of traffic safety, and there were mainly two methods to tackle the issue:

The first method that has been commonly used was data resampling, where the original dataset was resampled with under-sampling or over-sampling techniques. In other words, majority samples (e.g. less severe crashes) were discarded or new minority samples (e.g. more severe crashes) were generated. For instance, in recent years, data resampling methods have been utilised for crash severity modelling (Yahaya et al., 2021; Gupta et al., 2021; Rahimi et al., 2020; Fiorentini and Losa, 2020), driving assessment and risk prediction (Shi et al., 2019; Chen et al., 2020a), and real-time accident detection or prediction (Abou Elassad et al., 2020; Ke et al., 2019; Parsa et al., 2019; You et al., 2017). However, the data resampling method tends to distort the original distribution of the training data (Seiffert et al., 2008; Weiss et al., 2007), such that the minority class can be overfitted whereas the majority class can lose important information if discarded (Chawla, 2009; Liang and Zhang, 2012).

Another method to handle the class imbalance issue is cost-sensitive learning, which does not alter the original distribution of dataset. Cost-sensitive learning is a machine learning method that is able to alter the weights or misclassification cost assigned to various classes in a problem with class imbalance issue, such that fairer attention can be directed to the classification of the minority class (Liu and Zhou, 2006; Dabiri et al., 2020; Kuang et al., 2019). In the field of traffic safety, for instance, cost-sensitive learning methods have been applied for real-time crash prediction (Ke et al., 2019), accident duration prediction (Kuang et al., 2019), hit-and-run crash prediction (Zhu and Wan, 2021), and crash severity prediction (Iranitalab and Khattak, 2017), etc.

In this study, as the amount of AV crash records for analysis is limited, we would like to avoid data distortion thoroughly, hence the cost-sensitive learning method is adopted to mitigate the class imbalance issue in the AV crash dataset.

1.2. Objectives and contributions

The objective of this paper is to determine the most significant contributing factors to AV crash severity, based on the California AV collision database. To handle the class imbalance issue of AV crash dataset, the cost-sensitive CART approach is utilised to generate the classification tree. The effect of time, environment conditions, locations, vehicle characteristics, movements, driving modes, and collision types have been investigated.

The contributions of this study are two-fold. First, this paper investigates the contributing factors to AV crash severity, to which researchers have paid limited attention. Second, the relationship between AV crash severity and its influencing factors has been examined with the cost-sensitive CART approach, which could handle variable interaction and tackle the dataset's class imbalance issue, without relying on strict statistical assumptions. The graphical display of the tree structure makes it easy to understand the relationship between the AV crash severity and its influencing factors. The results could provide useful references for traffic engineers or AV manufacturers to raise effective counter measures or policies to mitigate AV crash severity.

The remainder of this paper is organised as follows: Section 2 introduces the data collection process and provides descriptive analysis of the dataset. Cost-sensitive CART model is formulated in Section 3. Section 4 shows the tree structure and its corresponding interpretations and discussions. Section 5 draws the conclusion, points out the current limitations and potential future research directions.

2. Data

In this section, we first present a review of the related datasets on AV, including both AV safety datasets and other autonomous driving datasets. Then, we describe the characteristics of AV crash records, which were extracted from the AV road test in California for the crash severity analysis in this study.

2.1. Review of datasets

After searching the data websites and reviewing the publications including both numerous research papers and review papers, the datasets on autonomous vehicles (AVs) have been summarised as follows:

2.1.1. AV safety datasets

Despite the continuous development of AV technology, the AV safety issue has attracted the attention of various stakeholders, including government agencies, manufacturers, and general public. Thereby, the AV safety related datasets have been overviewed and summarised in Table 1.

The California Department of Motor Vehicles' AV collision reports and disengagement reports are the most commonly utilised AV road test data in many recent research (Banerjee et al., 2018; Wang and Li, 2019a,c; Boggs et al., 2020a; Favarò et al., 2018), published by

Table 1
AV safety datasets.

No.	Data source	Description
1	California Department of Motor Vehicles (DMV, 2021)	AV collision & disengagement records
2	Manufacturer's self safety report	e.g. Google's safety report & safety white papers
3	General crash severity dataset	e.g. National Transportation Safety Board (NTSB, 2021) crash dataset

California Department of Motor Vehicles (DMV). With regard to the testing regulation on California roads, the AV manufacturers are necessitated to report their AV-involved collisions within 10 days. In addition, they are required to provide annual report on the frequency of their AVs disengaged from autonomous mode during tests. Moreover, AV manufacturers such as Google has provided general descriptive statistics on AV safety, etc. (Waymo, 2021a). General crash severity datasets (No. 3), such as National Transportation Safety Board (NTSB) crash dataset, are not dedicated to AV testing, even through they contain a large number of collision records, which could include AV collisions (Wang and Li, 2019c).

2.1.2. Autonomous driving datasets

In addition to the datasets on AV safety, we have also identified the AV/self-driving related datasets from other sources, where the datasets are collected by vehicle on-board sensors, and contain video camera, LiDAR or radar data. Guo et al. (2019) presented a recent comprehensive review of the AV related datasets prior to this study. These datasets have been examined and listed as follows, whereas eight datasets whose web links are inaccessible at present have been excluded here: Apollo Open Platform (Apollo, 2021a), Apolloscape (Apollo, 2021b); Berkeley DeepDrive dataset (BDD, 2021); BraineCar dataset including face camera and road camera data (Brain4Cars, 2021); Cambridge-driving Labeled Video Database (Brostow et al., 2008, 2009); CCSAD dataset (Guzmán et al., 2015); Cityscapes dataset (Cityscapes, 2021); Comma.ai driving dataset (Comma.ai, 2021); CULane dataset (Pan et al., 2018); DAVIS Driving Dataset (Binas et al., 2017); DBNet driving behaviour dataset (Chen et al., 2018); DIPLECS autonomous driving dataset (Pugeault and Bowden, 2015); DR(eye)VE dataset with drivers' gaze fixation information (Palazzi et al., 2018); EISATS dataset which contains image sequences for analysis (CVC research center, UAB and UPC universities, 2021); Elektra Autonomous Vehicle dataset (Reinhard Klette, 2021); Ford Campus Vision and Lidar Dataset (Pandey et al., 2011); HCI stereo and optical flow dataset (Meister et al., 2012); HD1K Benchmark Suite for optical flow (HD1K, 2021); Joint Attention in Autonomous Driving Dataset (Rasouli et al., 2017; Houben et al., 2013); KAIST Multi-spectral Day/Night dataset (Choi et al., 2018); KITTI Vision Benchmark Suite (Geiger et al., 2013); Mapillary Vistas imagery dataset (Neuhold et al., 2017); the Málaga Stereo and Laser Urban Dataset (Blanco et al., 2014); Oxford Robotcar dataset (Maddern et al., 2017); UAH-DriveSet (Romera et al., 2016); Udacity self-driving car dataset (Udacity, 2021). More details on various categorisation of these datasets by tasks such as stereo/3D vision and behavioural analysis can be found in Guo et al. (2019).

Specifically, there are abundant datasets which are solely dedicated to object detection or tracking, including Belgium traffic sign recognition dataset (Mathias et al., 2013), German Traffic Sign Recognition/Detection Benchmark datasets (Pugeault and Bowden, 2015; Houben et al., 2013) for traffic sign detection; Bosch Small Traffic Lights Dataset (Behrendt and Novak) for traffic light detection; Caltech Pedestrian Dataset (Caltech, 2021), Daimler Pedestrian Segmentation Benchmark Dataset (Gavrila, 2021), ETH Zurich robust multi-person tracking dataset (Ess et al., 2008), TUD-Brussels and TUD-MotionPairs dataset (Wojek et al., 2009) for pedestrian detection or tracking; highway image sequence dataset (Kondermann et al., 2014) for highway work zones detection; TME Motorway dataset (Caraffi et al., 2012) for vehicle detection and tracking; nuScenes dataset (Caesar et al., 2019), the Stanford Track Collection dataset (Teichman et al., 2011) for general urban scene object detection.

Furthermore, we have also enriched the list of datasets in Guo et al. (2019) by including 19 more publicly available recent AV related datasets. The data sources and brief descriptions of these datasets are summarised in Table 2, including images, videos, simulated data, other sensor data, and combinations of them.

2.2. Data collection and descriptive analysis

Based on the dataset review in the section above, the California Department of Motor Vehicles AV collision dataset (DMV, 2021), which provides abundant AV incident information on crash severity, is utilised in this study for crash severity analysis. According to the testing regulation on California roads, manufacturers are required to report the AV-involved collision within 10 days. 245 AV crash records from Oct-2018 to Oct-2021 have been collated. Table 3 shows the descriptive statistics of California AV crash dataset used in this study.

The target variable is the AV crash severity, which contains information about the human injuries in AV crashes. We have identified no fatal crashes and 6 cases with possible non-capacitating injury, where victims were transported to hospital or called emergency services. Due to the very limited number of fatal and serious injury (FSI) crashes (2%), FSI and other injury crashes are combined as one category, such that the development of a separate statistically significant functional relationship for FSI crash severity level is precluded. In other words, similar to Song et al. (2021), we classify the crash severity into two classes, i.e. injury (198 crashes) & no injury (47 crashes), where the class imbalance ratio is around 4.2:1. More detailed crash severity data was not available at the current stage, which could be analysed in the future work.

Table 4 further shows the distribution of AV collisions by AV crash severity and different variable categories, in which 'Injury' indicates the more severe crashes with injuries reported and 'No injury' indicates less severe crashes with no injuries reported. In the table, part (1) shows the distribution of AV manufacturers, in which Cruise, Waymo and Zoox are the most common ones. Part (2) depicts the distribution of AV collisions in different cities in California, where the majority of AV crashes occurred in San Francisco. Part (3) shows the distribution of AV driving modes, where the majority of AV was in autonomous mode when crash occurred. Part (4) gives the distribution of facility types. It is observed that the majority of AV crashes occur at intersections where conflicts among vehicles are more complex. Part (5) presents information about the movements of AV and second party preceding collision, as well as collision types: Proceeding straight and stopped were the two most common AV movements prior to collision; proceeding straight and changing lanes were the two most common second party movements before collision; rear end collision has the highest frequency among different types of collisions.

The original dataset contains 18 features and 245 AV crash records, from which we do data sparsification followed by near zero variance elimination (Kuhn et al., 2008) in order to remove the features that take constant or almost constant value across the entire crash dataset, as they are uninformative and could compromise the accuracy of the model. After data pre-processing, 51 binary crash contributing factors have been retained for further model formulation.

3. Research methodology

3.1. Classification and regression tree

The classification and regression tree (CART) model is a representative data mining method for predictive modelling problems. When the target variable is continuous, the regression tree is formulated, while

Table 2
Recent AV driving datasets.

No.	Data source	Description
1	Connected & Autonomous Transportation Systems Laboratory (CATS Lab) of University of South Florida (CATS, 2021)	Contains abundant open-source CAV GPS trajectory data, for instance, lane change data, platooning data, etc.
2	Audi Autonomous Driving Dataset (A2D2) (Geyer et al., 2020)	An Audi Q7 e-tron is equipped with multimodal sensor suite, and the driving data has been recorded and labelled
3	Argoverse Dataset (Chang et al., 2019)	AV test vehicles of Argo AI have been operated in Miami and Pittsburgh, providing 3D tracking annotations and vehicle trajectories
4	Leddar PixSet Dataset (Leddar, 2021)	Collects data from AV equipped with full sensor suite, containing full-wave form LiDAR data
5	Lyft Level 5 Open Data (Level 5, 2021)	Contains prediction dataset with movement data of different travel agents; perception dataset with raw camera and LiDAR data from AV fleet
6	PandaSet (Hesai and Scale AI, 2021)	Utilises LiDAR sensors and video cameras onboard AV, 28 annotation classes are indicated
7	Waymo Open Dataset (Waymo, 2021b)	Multimodal sensor dataset collected from Waymo AVs, including motion dataset and perception dataset
8	PreSIL Dataset (Hurl et al., 2019)	A Precise Synthetic Image and LiDAR (PreSIL) dataset for AV perception
9	Oxford Radar RobotCar Dataset (Barnes et al., 2020)	Utilises Millimetre-Wave FMCW scanning radar data, serves as a radar extension to the previous Oxford Robotcar dataset (Maddern et al., 2017)
10	Drive&Act Dataset (Martin et al., 2019)	Contains video data for AV drivers' behaviour recognition
11	RUGD dataset (Wigness et al., 2019)	Contains video sequences, emphasising on understanding the application of the unstructured outdoor environments in off-road autonomous navigation
12	DrivingStereo dataset (Yang et al., 2019)	Contains images covering various sets of driving scenarios
13	Italdesign dataset (IDDA) (Alberti et al., 2020)	A large-scale multi-domain dataset for autonomous driving under different weather conditions, view point conditions, city types
14	COMAP dataset (Yuan and Sester, 2021)	A co-simulation synthetic autonomous driving perception dataset based on CARLA and SUMO simulator
15	Indian Driving dataset (IDD) (Varma et al., 2019)	Contains images collected from a front facing camera fixed on moving car in India
16	Test Area Autonomous Driving Baden-Württemberg Dataset (TAF-BW Dataset) (Zipfl et al., 2020)	Recorded by the intelligent and connected infrastructure, contains trajectory data, traffic signal information and georeferenced maps
17	One Million Scenes (ONCE) dataset (Mao et al., 2021)	Contains LiDAR data and images for 3D object detection in AV navigation
18	WoodScape dataset (Yogamani et al., 2019)	Fisheye dataset with images collected by four surrounded cameras for autonomous driving
19	Brno Urban Dataset (Ligocki et al., 2020)	A navigation and localisation dataset for AVs, collected by cameras, infrared camera, inertial measurement unit, etc.

a classification tree is developed when the target variable is discrete. Since this paper is concerned with the AV crash severity level, which is a categorical variable, a classification tree is modelled. The classification tree building algorithm is illustrated in the following steps. In

Table 3
Descriptive statistics of California AV crash dataset.

Variable	Description	Count	%
Severity	Injury	47	19%
	No Injury	198	81%
Month	Jan	21	9%
	Feb	22	9%
	Mar	18	7%
	Apr	11	4%
	May	17	7%
	Jun	21	9%
	Jul	27	11%
	Aug	28	11%
	Sep	16	7%
	Oct	28	11%
	Nov	22	9%
	Dec	14	6%
Year	2018	23	9%
	2019	105	43%
	2020	44	18%
	2021	73	30%
Time	Morning peak	24	10%
	Evening peak	19	8%
	Non-peak	201	82%
Manufacturer	AImotive	1	1%
	Apple	4	2%
	Argo	2	1%
	Aurora	3	1%
	Cruise	107	44%
	Lyft	8	3%
	Pony.ai	5	2%
	Waymo	85	35%
	Weride	2	1%
	Zoox	28	11%
AV was	Moving	139	57%
	Parked	4	2%
	Stopped in traffic	102	42%
Number of vehicles involved	1	35	14%
	2	208	85%
	3	2	1%
City	Palo Alto	13	5%
	Fremont	1	1%
	Irvine	2	1%
	Los Altos	3	1%
	Los Angeles	1	1%
	Milpitas	2	1%
	Mountain View	21	9%
	San Francisco	182	74%
	Santa Clara	19	8%
	Sunnyvale	1	1%
2nd party type	Vehicle	206	84%
	Vulnerable road user	28	11%
	Object	11	4%
2nd party was	Moving	212	87%
	Parked	16	7%
	Stopped in traffic	6	2%
Driving mode	Autonomous	146	60%
	Conventional	99	40%
Facility type	Intersection	164	67%
	Parking lot	18	7%
	Road segment	63	26%
Weather condition	Clear	214	87%
	Cloudy	22	9%
	Raining	6	2%
	Fog/visibility	3	1%
Lighting condition	Daylight	180	73%
	Dark with street lights	57	23%
	Dusk–dawn	8	3%
Road surface	Dry	222	91%
	Wet	13	5%
Roadway conditions	No unusual condition	215	88%
	Construction-repair zone	2	1%
	Reduced roadway width	3	1%
	Obstruction on roadway	1	1%
AV movement pre-	Backing	7	3%

(*continued on next page*)

Table 3 (*continued*).

Variable	Description	Count	%
ceding collision	Changing lanes	7	3%
	Entering traffic	1	1%
	Making left turn	18	7%
	Making right turn	17	7%
	Merging	3	1%
	Parked	6	2%
	Parking manoeuvre	4	2%
	Passing other vehicle	2	1%
	Proceeding straight	69	28%
	Slowing/stopping	21	9%
	Stopped	90	37%
2nd party movement preceding collision	Backing	8	3%
	Changing lanes	23	9%
	Entering traffic	5	2%
	Making left turn	12	5%
	Making right turn	18	4%
	Other unsafe turning	5	2%
	Parked	15	6%
	Parking manoeuvre	8	3%
	Passing other vehicle	9	4%
	Proceeding straight	114	47%
	Slowing/stopping	4	2%
	Stopped	5	2%
	Crossing into opposing lane	2	1%
Collision type	Broadside	23	9%
	Head on	16	7%
	Hit object	11	4%
	Rear end	135	55%
	Sideswipe	51	21%
	Passing other vehicle	1	1%
	Others	8	3%

the first step, the algorithm aims to build the tree. Let N be a parent node, and N_L and N_R be its two child nodes. The target variable–AV crash severity is recursively partitioned in order to maximise the purity level or impurity reduction $\Delta G(N)$ at each node N, which is calculated by subtracting the weighted impurities from the child nodes from the original impurity of the parent node:

$$\Delta G(N) = G(N) - \frac{m(N_L)}{m(N)} \cdot G(N_L) - \frac{m(N_R)}{m(N)} \cdot G(N_R) \tag{1}$$

where $m(N)$, $m(N_L)$, $m(N_R)$ represent the number of AV crash records at nodes N, N_L and N_R respectively. In this study, the impurity measure or function G is defined as the Gini index:

$$G(N) = 1 - \sum_j [p(j|N)]^2 \tag{2}$$

where j represents the class of target variable, $j \in \{1, \ldots, J\}$. In this study, j refers to the AV crash severity level. When the value of $G(N)$ is zero, all crash records in the node belong to a single group of AV crash severity level, demonstrating the least impurity. The probability of an AV crash record belonging to class j provided that it exists in node N is denoted by $p(j|N)$, which can be calculated as:

$$p(j|N) = \frac{p(j, N)}{p(N)} \tag{3}$$

where the numerator and denominator of the right-hand-side of the equation can be calculated with Eqs. (4) and (5), respectively:

$$p(j, N) = \frac{\pi(j) \cdot m_j(N)}{m_j} \tag{4}$$

where $p(j, N)$ denotes the probability of AV crash is both with class j and in node N; $\pi(j)$ denotes the prior probability of class j, i.e., the probability that an AV crash with class j severity level is presented to the tree; $m_j(N)$ represents the number of AV crashes with class j severity level presented at node N; m_j represents the number of class j crashes presented in the tree.

$$p(N) = \sum_j p(j, N) \tag{5}$$

Table 4
Distribution of AV collisions by AV crash severity & categories.

Feature	Injury	%	No injury	%
(1) Distribution of AV manufacturers				
AImotive	0	0%	1	100%
Apple	0	0%	4	100%
Argo	1	50%	1	50%
Aurora	0	0%	3	100%
Cruise	33	31%	74	69%
Lyft	0	0%	8	100%
Pony.ai	0	0%	5	100%
Waymo	12	14%	73	86%
Weride	0	0%	2	100%
Zoox	1	4%	27	96%
(2) Distribution of cities				
Palo Alto	2	15%	11	85%
Fremont	0	0%	1	100%
Irvine	0	0%	2	100%
Los Altos	0	0%	3	100%
Los Angeles	0	0%	1	100%
Milpitas	0	0%	2	100%
Mountain View	3	17%	18	83%
San Francisco	40	22%	142	78%
Santa Clara	2	11%	17	89%
Sunnyvale	0	0%	1	100%
(3) Distribution of driving modes				
Autonomous	32	22%	114	78%
Conventional	15	15%	84	85%
(4) Distribution of facility types				
Intersection	40	32%	124	68%
Parking lot	2	11%	16	89%
Road segment	5	8%	58	92%
(5) Top 5 movements and collision types				
AV's movement preceding collision				
Stopped	15	17%	75	83%
Proceeding straight	12	17%	57	83%
Slowing/stopping	6	29%	15	71%
Making left turn	7	39%	11	61%
Making right turn	5	29%	12	71%
2nd party's movement preceding collision				
Proceeding straight	33	29%	81	71%
Changing lanes	3	13%	20	87%
Parked	0	0%	15	100%
Making right turn	5	28%	13	72%
Making left turn	1	8%	11	92%
Collision types				
Rear end	32	24%	103	76%
Sideswipe	3	6%	48	94%
Broadside	7	30%	16	70%
Head on	3	19%	13	81%
Hit object	0	0%	11	100%

Note: % – row percentages across severity categories.

where $p(N)$ denotes the probability of any AV crash that falls into node N.

After recursively partitioning the data to maximise the level of purity in each leaf node, a saturated classification tree is obtained. However, the saturated tree could result in overfitting, which does not help in the classification of the validation dataset. Therefore, in the second step, the classification tree is pruned. In other words, cross-validation is applied to trim back the saturated tree, such that simpler tree with better cross-validated performance can be obtained (Therneau et al., 2019).

3.2. Cost-sensitive CART model

In this study, the distribution of the AV crash severity dataset is class-imbalanced, as the number of crashes with injuries reported is much less than that without any injuries reported. Thereby, the

classification result will be biased toward the majority class with less severe AV crash such that the classification model's performance can be compromised. To handle the class imbalance issue in the classification tree model, the cost-sensitive learning process can be embedded into the traditional CART model. A misclassification cost matrix C of size $J \times J$ can be specified, where the column indicates the actual class of AV crash severity, the row indicates the predicted class, and $c(i,j)$ indicates the cost of misclassifying j as i. Thereby, C has zero on the diagonal ($c(i,j) = 0,\ \forall i = j$), and positive values off diagonal ($c(i,j) > 0,\ \forall i \neq j$) to indicate the cost of misclassification. In the cost-insensitive classification tree, the off-diagonal values are set to 1 ($c(i,j) = 1,\ \forall i \neq j$), such that the importance of misclassifying various classes is treated equally and the prior probabilities are equal to the observed class frequencies in the dataset. On the other hand, the cost-sensitive learning process adjusts the biased classification model by allocating a higher penalty value for misclassifying the sample from the minority class (Ke et al., 2019). At least one of the $c(i,j)$ value can be set as 1 without loss of generality (Ting, 1998).

In particular, the misclassification cost can be incorporated into the cost-sensitive decision tree induction through altered priors $\pi'(y)$. In other words, the prior probability value $\pi(y)$ in Eq. (4) is replaced with the Eq. (6):

$$\pi'(j) = \frac{\pi(j)\cdot C(j)}{\sum_i \pi(i)\cdot C(i)} = \frac{m_j \cdot C(j)}{\sum_i m_i \cdot C(i)} \tag{6}$$

where

$$C(j) = \sum_i cost(i,j) \tag{7}$$

$$\pi(j) = \frac{m_j}{m},\ \forall j \in \{1,\ldots,J\} \tag{8}$$

where $\pi(j)$ represents the priori, m denotes the total number of AV crashes. With Eq. (6), by putting a larger priori on the minority class, its misclassification rate tends to reduce. More details on the induction process of cost-sensitive decision tree algorithm can be referred to Breiman et al. (1984). When altered priors are utilised, only the choice of split would be affected (Therneau et al., 2019). In other words, the altered priors will assist the impurity rule in choosing splits via Eq. (4).

3.3. Other comparative methods

In this study, to compare with the performance of cost-sensitive/traditional CART model, we have also applied other machine learning/statistical models, including random forest, boosted classification tree, artificial neural network and logistic regression, which are illustrated as follows:

Both random forest model and boosted classification tree are ensemble learning algorithms, such that many based learners or trees have been fitted and predictors are aggregated across them. For the boosted classification tree, each base learner learns from the mistakes of the fits in the previous iteration. More details on the boosting trees can be found in Hastie et al. (2009), Kuhn et al. (2013). On the other hand, instead of generating base tree models in sequence like the boosted classification tree algorithm, the base learners or individual trees in random forest are modelled with bootstrapped samples of data in parallel way. At splitting nodes of the individual trees, only a random subset of features are utilised, hence the algorithm enables the diversity among base learners (Breiman, 2001; Zhang and Haghani, 2015).

The artificial neural network algorithm, which contains an assembly of inter-connected nodes (artificial neurons) and weighted links, attempts to model the way that human brains with a network of neurons learning and extracting information (Schmidhuber, 2015). In this study, the input neurons receive the AV crash data, which is further passed to the neurons in the hidden layer through weighted connections, which indicate their relative importance. The weighted sum of the inputs is then applied to activation function to obtain the response (Kumar et al., 2015), i.e. AV crash severity classification. More details can be referred to Ripley (2007).

		Actual label: Positive	Actual label: Negative
Predicted label	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Fig. 1. Confusion matrix.

Lastly, the logistic regression is a statistical model which utilises the logistic function (as shown in Eq. (9)) to model the probability of various classes of dependent variable (Kleinbaum and Klein, 2010), and it has been commonly used in the literature regarding crash severity analysis (Decker et al., 2016; Ahmadi et al., 2020; Salon and McIntyre, 2018; Yan et al., 2011).

$$p = \frac{e^{(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}{1+e^{(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \tag{9}$$

For the binary AV crash severity classification problem in this study, the response variable is represented by a dummy variable (1 indicates more severe crash and 0 indicates less severe crash); and other crash contributing factors serve as the explanatory variables.

4. Results and discussions

In this section, we demonstrate the classification tree structure generated by the cost-sensitive CART model and provide related discussions. Cost-insensitive classification tree is also constructed for comparison and the effect of class-weighted ratio (c_+/c_-) has been illustrated.

4.1. Experimental settings

The proposed AV crash severity modelling problem is basically a binary classification task. According to the confusion matrix shown in Fig. 1, we set the minority class j with injuries reported as the positive class, and the crashes with no injuries reported i as the negative class. To handle the class imbalance issue, the value of $c_+ = c(i,j)$ should be more expensive than $c_- = c(j,i)$. In other words, the cost values c_+ and c_- control the balance between more severe and less severe AV crashes respectively.

For performance evaluation of cost-sensitive classification tree model, four commonly used evaluation metrics are adopted, including the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity and specificity.

$$Sensitivity = \frac{TP}{TP+FN} \tag{10}$$

$$Specificity = \frac{TN}{TN+FP} \tag{11}$$

where true positive (TP), false positive (FP), true negative (TN) and false negative (FP) values are demonstrated with Fig. 1. Note that in the binary classification scenario in this study, the value of sensitivity and specificity can indicate the performance of the models in predicting more severe and less severe AV crashes respectively. When the number of AV crash severity levels is three and above, the modelling process becomes a multi-class classification task, where the performance of each AV crash severity level should be computed and analysed separately.

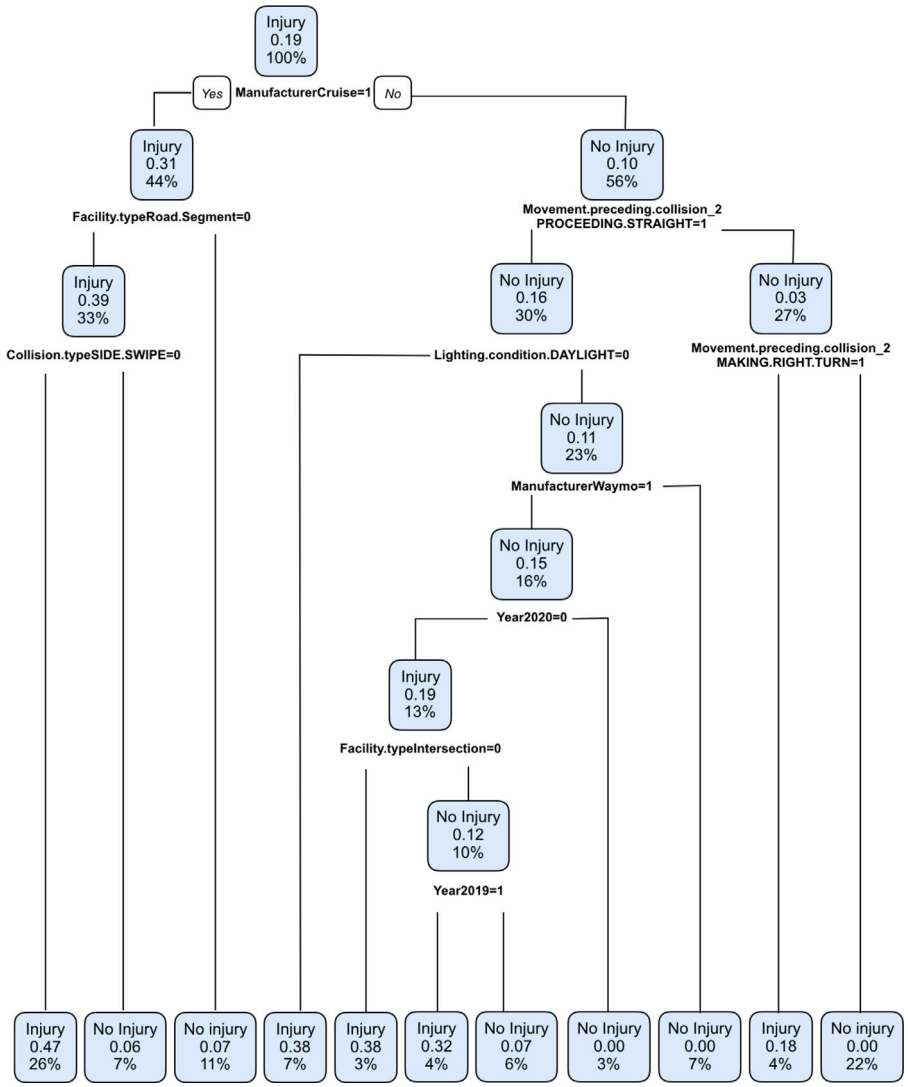

Fig. 2. Classification tree structure.

4.2. Tree structure interpretation

With ten-fold cross validation, we have tuned the complexity parameter in the cost-sensitive classification tree to control and optimise the tree size. The optimal AUC value is 0.74 when the value of complexity parameter is 0.001 and the value of class-weighted ratio is 5. The corresponding cost-sensitive classification tree structure of AV crash severity is depicted in Fig. 2. From the root node to leaf nodes, all AV crashes with injury reported or no injury reported can be divided into 20 subgroups through various tree branches. The predicted class (whether injury has been reported or not, i.e. more or less severe crash), the probability of AV crash severity with injuries reported, as well as the percentage of observations (AV crashes) have been articulated corresponding to each node from top to bottom in Fig. 2. In the leaf nodes, the probability of AV crashes with injury reported ranges between 0 to 0.47. Different splitting variables are utilised in the tree structure, including manufacturer, facility type, movement preceding collision, collision type, light condition and year.

Interpretation of the graphical tree structure can be provided as follows: The first optimal splitting variable in the tree structure is manufacturer (Cruise). In the group of crashes involving vehicles manufactured by Cruise, we find that 31% crashes fallen in this category are reported with injury, and 44% of crashes fall into this rule. On the other hand, when other brands of AVs are involved, the probability of crashes with injury reported is 10%, covering 56% of the crashes.

In the second level of the tree structure, the facility type (road segment) leads to the split in the group of crashes with AVs manufactured by Cruise into two categories. The probability of more severe AV crash is 0.07 among crashes which occurs at road segment, and 0.39 among crashes which does not occur at road segment respectively. The result could be explained by the less complex vehicle movements and conflicts at road segments (Xu et al., 2019). Also in the second level of the tree structure, for the sub-group of crashes where the AV manufacturers are not Cruise, when second party of collision is proceeding straight before collision, the probability of more severe AV crash is 0.16, which is higher than that of the other branch, which could be explained by the high level of kinetic energy.

In the third level of decision tree, collision type (side swipe) segments the crashes which occur at road segment into two groups. When the collision type is side swipe, the likelihood of more severe crash is only 0.06, covering 7% of all crashes. On the other hand, when the collision type is not side swipe, the probability of more severe AV crash is higher as 0.47, covering 26% of all crashes. The result is consistent with that of human-driving vehicles (HVs) (Cerwick et al., 2014; Theofilatos et al., 2021), which could be explained by the angle and relative speed between the conflict parties. Also in the third level of the tree, when the movement preceding collision of the second party

is making right turn, the probability of severer crash tends to be higher as 0.18. The result can be explained by the challenge in AV's vision-based navigation to avoid obstacles around urban corners with various radii (Hubschneider et al., 2017; Cai et al., 2019).

In addition, in the third level of the tree structure, when the movement preceding collision of the second party is proceeding straight, if the light condition is daylight, the likelihood of more severe crashes is lower as 0.11, covering 16% of AV crashes. The result is consistent with the literature (Tarmizi and Abd Aziz, 2018; Wiseman et al., 2021; Wang et al., 2019), which suggested that AVs could perform road users detection more successfully during day time than the night scene or poor lighting condition. Thereby, intensive research works such as (Pham and Yoo, 2020; Gan et al., 2019) have been conducted in recent years to ensure the AV navigation safety under various lighting conditions. In the fourth level of the tree, when the light condition is daylight, the AV which is not manufactured by Waymo tends to be involved in less severe AV crash, as all crashes fallen in this category are with no injuries reported, covering 7% of the crashes. Next, in the fifth level of the tree, when the AV is manufactured by Waymo, crashes occurred in 2020 had no injuries reported, covering 3% of the AV crashes. Furthermore, we proceed to the sixth level of the tree, where the facility type intersection leads to the split in the group of crashes happened out of 2020. The probability of more severe AV crash is 0.12 among crashes occurred at intersections and 0.38 among crashes not occurred at intersections respectively. The result can possibly be attributed to the interaction effect between good detection visibility under daylight condition and AV's control system, which could be conservative for intersection control in order to avoid potential conflicts and risks (Noh, 2018; Sezer et al., 2015). Lastly, in the seventh level of decision tree, crashes occurred in 2019 are likely to be more severe, which could be explained by the less advanced AV control technique in the past.

4.3. Merits of the cost-sensitive CART model

We have compared the performance of the cost-sensitive classification tree model with the corresponding cost-insensitive model. The effects of different class-weighted ratios on AUC, accuracy, sensitivity, and specificity have been analysed, as shown Fig. 3. The values in the horizontal axis indicate the class-weighted ratios γ:

$$\gamma = c_+/c_- \tag{12}$$

where γ indicates the ratio of the cost of misclassifying crashes with injuries reported (c_+) to that of misclassifying crashes without injuries reported (c_-). By fixing $c_- = 1$ in the experiment, γ can indicate the importance of the positive class, i.e., crashes with injuries reported.

Observing Fig. 3, when the cost-sensitive learning is not taken into consideration for model development ($\gamma = 1$), AUC value is around 0.71 (as shown in Fig. 3(a)). On the other hand, Fig. 3(b) shows that high accuracy value can be achieved when $\gamma = 1$, which could be explained by the highest value of specificity which is 0.94 (as shown in Fig. 3(d)). However, the sensitivity value (as shown in Fig. 3(c)) is 0.23, which means that few of the more severe AV crashes have been correctly classified. Generally, the trade-off between sensitivity and specificity value can be demonstrated with Figs. 3(c) and 3(d). Sensitivity value can indicate the accuracy in terms of classifying crashes with more severe AV crashes. On the other hand, specificity can represent the accuracy in classifying less severe AV crashes. When $\gamma = 1$, the result suffices to show that the cost-insensitive model is biased toward the majority crashes without injuries reported. From Fig. 3(a), it is also observed that the highest AUC value 0.74 is given by $\gamma = 5$, demonstrating a good separability between more severe and less severe AV crashes. In comparison with the cost-insensitive model, the cost-sensitive decision tree sacrifices the performance in accuracy and specificity such that the sensitivity value can be improved.

Table 5
Models for comparison.

Method	AUC	Accuracy	G-mean	Sensitivity	Specificity
Cost-sensitive CART	0.74	0.66	0.67	0.69	0.65
CART	0.71	0.81	0.46	0.23	0.94
Random forest	0.64	0.78	0.45	0.22	0.92
Boosted classification tree	0.66	0.80	0	0	1
Artificial neural network	0.69	0.77	0.54	0.34	0.87
Logistic regression	0.67	0.85	0.60	0.38	0.96

4.4. Comparison with other models

Furthermore, we have also compared the performance of cost-sensitive CART model and CART model with other machine learning methods (random forest, artificial neural network, boosted classification tree) and logistic regression model. Note that even though ordinal regression technique has been utilised in previous AV crash severity analysis work (Wang and Li, 2019b), the number of FSI crashes in this paper is too few to generate statistically significant classification result. Therefore, logistic regression model instead of ordered logit model is adopted for benchmarking.

The results of model comparison have been summarised in Table 5. It can be observed that the advanced machine learning methods suffer from the class-imbalance issue, biasing toward higher specificity value, even though satisfiable accuracy value can be obtained. Note that due to the skewed class distribution in this study, we cannot merely utilise accuracy as the performance evaluation metric, as it could not distinguish between the accuracy in classifying various classes. In terms of AUC value and the G-mean of sensitivity and specificity, the cost-sensitive CART model outperforms the benchmarking methods as it tends to value the classification of minority AV crashes where injuries have been reported. On the other hand, the overall accuracy of cost-sensitive CART model is the worst, as it sacrifices the classification performance of the majority class. This study is limited to the cost-sensitive model of CART. More robust cost-sensitive classifiers could be explored and developed based on more advanced machine learning methods in the future work.

5. Conclusions

This paper investigated the effects of time, environment conditions, locations, vehicle characteristics, movements, driving modes, and collision types on AV crash severity using the classification tree approach based on the California AV road test data. In addition, the AV crash dataset suffers from the class imbalance issue which compromises the performance of AV crash severity classification model, as the number of more severe crashes is relatively few. Thereby, the cost-sensitive classification and regression tree (CART) model has been applied to investigate the rationale behind the occurrences of more severe crashes. A classification tree with 20 subgroups has been constructed to graphically demonstrate the relationship between AV crash severity and its influencing factors.

The results of the classification tree shows that the AV crash severity level is associated with manufacturer, facility type, movement preceding collision, collision type, light condition and year. Sensitivity analysis has also been conducted to analyse the effect of the class-weighted ratio, which demonstrates that the cost-sensitive CART model outperforms the CART model with unified misclassification cost. The experimental results have also been compared with other machine learning and statistical methods, where better performance of cost-sensitive CART model has been demonstrated, as the class imbalance issue has been tackled specifically.

Based on the results of the paper, the study specific recommendations to traffic engineers and AV manufacturers on AV crash severity mitigation and safety improvement are discussed as follows: AV manufactured by Cruise and Waymo are more likely to be involved in more

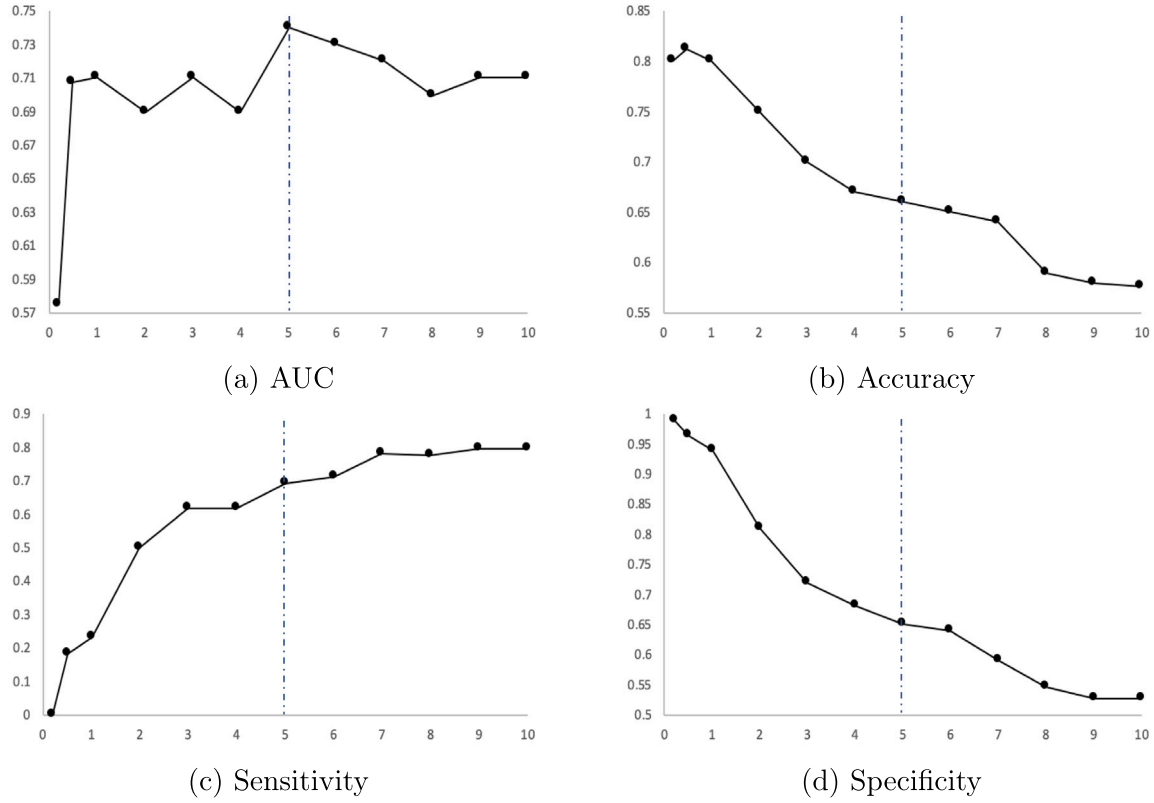

Fig. 3. Class imbalance cost.

severe crashes, and these manufacturers could review the accident record carefully to avoid potential fault as the side of AV control and detection system, as well as the AV operator. The AV crash severity tends to be lower when the crash location is road segment as the traffic situation is not complex. On the other hand, intersections have been identified as the hotspots of AV crashes, hence it is recommended to explore more advanced AV intersection control algorithms to mitigate the potential conflicts. The crash severity level is likely to be more severe when the opponent vehicle is proceeding straight and making right turn. On the other hand, similar to the studies on HVs, AV crash tends to be less severe when the collision type is sideswipe. Thereby, the AV operator is also suggested to be alert when the opponent vehicle is proceeding straight or making right turn due to kinetic energy and visibility concern, meanwhile, the more advanced road user detection system could be investigated by traffic engineers. In addition, under daylight condition, the likelihood of severe AV crashes tends to be lower, whereas the AV detection system under poor light condition is recommended to be improved, which could draw the attention of researchers and manufacturers in the future.

This study is limited to the information provided by the current AV crash dataset for analysis and the data size. In the future, the methodology could be extended to include more crash records and AV crash-related features from available data sources. For instance, more detailed video camera and sensor data analysis could be conducted to explore more detailed AV-specific crash information. More systematic analysis should be conducted when autonomous vehicles are more widely used and the safety-related data are more complete. Furthermore, more advanced cost-sensitive learning algorithms could also developed and investigated based on state-of-art machine learning techniques.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study is supported by the Ministry of Education of Singapore, via the project R-302-000-286-114 under the MOE Tier 1 Grant FY2021. Any opinions, findings and conclusions or recommendations expressed in this study are those of the author(s) and do not reflect the views of the Ministry of Education of Singapore.

References

Abou Elassad, Z.E., Mousannif, H., Al Moatassime, H., 2020. A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems. Transp. Res. C 118, 102708.

Ahmadi, A., Jahangiri, A., Berardi, V., Machiani, S.G., 2020. Crash severity analysis of rear-end crashes in california using statistical and machine learning classification methods. J. Transp. Saf. Secur. 12 (4), 522–546.

Alambeigi, H., McDonald, A.D., Tankasala, S.R., 2020. Crash themes in automated vehicles: A topic modeling analysis of the california department of motor vehicles automated vehicle crash database. arXiv preprint arXiv:2001.11087.

Alberti, E., Tavera, A., Masone, C., Caputo, B., 2020. IDDA: A large-scale multi-domain dataset for autonomous driving. IEEE Robot. Autom. Lett. 5 (4), 5526–5533.

Apollo, 2021a. Apollo autonomous vehicle dataset. https://apollo.auto/southbay.html, Online; accessed 12 Dec 2021.

Apollo, 2021b. Apolloscape dataset. http://apolloscape.auto/scene.html, Online; accessed 12 Dec 2021.

Banerjee, S.S., Jha, S., Cyriac, J., Kalbarczyk, Z.T., Iyer, R.K., 2018. Hands off the wheel in autonomous vehicles?: A systems perspective on over a million miles of field data. In: 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, pp. 586–597.

Barnes, D., Gadd, M., Murcutt, P., Newman, P., Posner, I., 2020. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 6433–6438.

BDD, 2021. Berkeley DeepDrive dataset. https://bdd-data.berkeley.edu, Online; accessed 14 Dec 2021.

Behrendt, K., Novak, L., A deep learning approach to traffic lights: Detection, tracking, and classification. In: Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE.

Binas, J., Neil, D., Liu, S.-C., Delbruck, T., 2017. DDD17: End-to-end DAVIS driving dataset. arXiv preprint arXiv:1711.01458.

Blanco, J.L., Moreno, F.A., Gonzalez-Jimenez, J., 2014. The málaga urban dataset: High-rate stereo and lidars in a realistic urban scenario. Int. J. Robot. Res. 33 (2), 207–214.

Boggs, A.M., Arvin, R., Khattak, A.J., 2020a. Exploring the who, what, when, where, and why of automated vehicle disengagements. Accid. Anal. Prev. 136, 105406.

Boggs, A.M., Wali, B., Khattak, A.J., 2020b. Exploratory analysis of automated vehicle crashes in california: A text analytics & hierarchical Bayesian heterogeneity-based approach. Accid. Anal. Prev. 135, 105354.

Brain4Cars, 2021. How does Brain4Cars work?. http://brain4cars.com, Online; accessed 14 Dec 2021.

Breiman, L., 2001. Random forests. Mach. :Earn. 45 (1), 5–32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Brostow, G.J., Fauqueur, J., Cipolla, R., 2009. Semantic object classes in video: A high-definition ground truth database. Pattern Recognit. Lett. 30 (2), 88–97.

Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R., 2008. Segmentation and recognition using structure from motion point clouds. In: European Conference on Computer Vision. Springer, pp. 44–57.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2019. Nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027.

Cai, P., Sun, Y., Chen, Y., Liu, M., 2019. Vision-based trajectory planning via imitation learning for autonomous vehicles. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, pp. 2736–2742.

Caltech, 2021. Caltech pedestrian detection benchmark. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/, Online; accessed 14 Dec 2021.

Caraffi, C., Vojíř, T., Trefnỳ, J., Šochman, J., Matas, J., 2012. A system for real-time detection and tracking of vehicles from a single car-mounted camera. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems. IEEE, pp. 975–982.

CATS, 2021. Connected & autonomous transportation systems. https://github.com/CATS-Lab-USF?tab=overview&from=2021-12-01&to=2021-12-10, Online; accessed 10 Dec 2021.

Cerwick, D.M., Gkritza, K., Shaheed, M.S., Hans, Z., 2014. A comparison of the mixed logit and latent class methods for crash severity analysis. Anal. Methods Accid. Res. 3, 11–27.

Chang, L.Y., Chien, J.T., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Saf. Sci. 51 (1), 17–22.

Chang, M.F., Lambert, J.W., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J., 2019. Argoverse: 3D tracking and forecasting with rich maps. In: Conference on Computer Vision and Pattern Recognition (CVPR).

Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accid. Anal. Prev. 38 (5), 1019–1027.

Chawla, N.V., 2009. Data mining for imbalanced datasets: An overview. Data Min. Knowl. Discov. Handb. 875–886.

Chen, T., Shi, X., Wong, Y.D., Yu, X., 2020a. Predicting lane-changing risk level based on vehicles' space-series features: A pre-emptive learning approach. Transp. Res. C 116, 102646.

Chen, Y., Wang, J., Li, J., Lu, C., Luo, Z., Xue, H., Wang, C., 2018. Lidar-video driving dataset: Learning driving policies effectively. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5870–5878.

Chen, S., Wang, H., Meng, Q., 2019. Designing autonomous vehicle incentive program with uncertain vehicle purchase price. Transp. Res. C 103, 226–245.

Chen, S., Wang, H., Meng, Q., 2020b. Solving the first-mile ridesharing problem using autonomous vehicles. Comput.-Aided Civ. Infrastruct. Eng. 35 (1), 45–60.

Chen, S., Wang, H., Meng, Q., 2021a. Autonomous truck scheduling for container transshipment between two seaport terminals considering platooning and speed optimization. Transp. Res. B 154, 289–315.

Chen, S., Wang, H., Meng, Q., 2021b. An optimal dynamic lane reversal and traffic control strategy for autonomous vehicles. IEEE Trans. Intell. Transp. Syst..

Chen, S., Wang, H., Xiao, L., Meng, Q., 2022. Random capacity for a single lane with mixed autonomous and human-driven vehicles: Bounds, mean gaps and probability distributions. Transp. Res. Part E: Logist. Transp. Rev. 160, 102650.

Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J.S., An, K., Kweon, I.S., 2018. KAIST multi-spectral day/night data set for autonomous and assisted driving. IEEE Trans. Intell. Transp. Syst. 19 (3), 934–948.

Cityscapes, 2021. Cityscapes dataset. http://research.comma.ai, Online; accessed 13 Dec 2021.

Comma.ai, 2021. Comma.ai driving dataset. https://www.cityscapes-dataset.com, Online; accessed 13 Dec 2021.

CVC research center, UAB, UPC universities, 2021. Elektra. http://adas.cvc.uab.es/elektra/datasets/, Online; accessed 14 Dec 2021.

Dabiri, S., Marković, N., Heaslip, K., Reddy, C.K., 2020. A deep convolutional neural network based approach for vehicle classification using large-scale GPS trajectory data. Transp. Res. C 116, 102644.

Das, S., Dutta, A., Tsapakis, I., 2020. Automated vehicle collisions in california: Applying Bayesian latent class model. IATSS Res. 44 (4), 300–308.

Decker, S., Otte, D., Cruz, D.L., Müller, C.W., Omar, M., Krettek, C., Brand, S., 2016. Injury severity of pedestrians, bicyclists and motorcyclists resulting from crashes with reversing cars. Accid. Anal. Prev. 94, 46–51.

Dixit, V.V., Chand, S., Nair, D.J., 2016. Autonomous vehicles: Disengagements, accidents and reaction times. PLoS One 11 (12), e0168054.

DMV, 2021. Autonomous vehicle collision reports. https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/, Online; accessed 05 May 2021.

Ess, A., Leibe, B., Schindler, K., van Gool, L., 2008. A mobile vision system for robust multi-person tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08). IEEE Press.

Favarò, F., Eurich, S., Nader, N., 2018. Autonomous vehicles' disengagements: Trends, triggers, and regulatory limitations. Accid. Anal. Prev. 110, 136–148.

Favarò, F.M., Nader, N., Eurich, S.O., Tripp, M., Varadaraju, N., 2017. Examining accident reports involving autonomous vehicles in california. PLoS One 12 (9), e0184952.

Fiorentini, N., Losa, M., 2020. Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures 5 (7), 61.

Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A., 2019. Self-supervised moving vehicle tracking with stereo sound. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7053–7062.

Gavrila, 2021. Daimler pedestrian segmentation benchmark dataset. http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/daimler_pedestrian_benchmark_d.html, Online; accessed 11 Dec 2021.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The kitti dataset. Int. J. Robot. Res. 32 (11), 1231–1237.

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., Schuberth, P., 2020. A2D2: Audi autonomous driving dataset. arXiv:2004.06320.

Goodall, N.J., 2021. Comparison of automated vehicle struck-from-behind crash rates with national rates using naturalistic data. Accid. Anal. Prev. 154, 106056.

Guo, J., Kurup, U., Shah, M., 2019. Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving. IEEE Trans. Intell. Transp. Syst. 21 (8), 3135–3151.

Gupta, R., Asgari, H., Azimi, G., Rahimi, A., Jin, X., 2021. Analysis of Truck-Involved Work Zone Crash Fatalities in Florida. Technical Report.

Guzmán, R., Hayet, J.-B., Klette, R., 2015. Towards ubiquitous autonomous driving: The CCSAD dataset. In: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 582–593.

Han, J., Pei, J., Kamber, M., 2011. Data Mining: Concepts and Techniques. Elsevier.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Boosting and additive trees. In: The Elements of Statistical Learning. Springer, pp. 337–387.

HD1K, 2021. HD1K benchmark suite. http://hci-benchmark.iwr.uni-heidelberg.de, Online; accessed 13 Dec 2021.

Hesai and Scale AI, 2021. Pandaset. https://pandaset.org, Online; accessed 12 Dec 2021.

Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C., 2013. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In: International Joint Conference on Neural Networks. (1288).

Huang, H., Peng, Y., Wang, J., Luo, Q., Li, X., 2018. Interactive risk analysis on crash injury severity at a mountainous freeway with tunnel groups in China. Accid. Anal. Prev. 111, 56–62.

Hubschneider, C., Bauer, A., Weber, M., Zöllner, J.M., 2017. Adding navigation to the equation: Turning decisions for end-to-end vehicle control. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 1–8.

Hurl, B., Czarnecki, K., Waslander, S., 2019. Precise synthetic image and lidar (presil) dataset for autonomous vehicle perception. In: 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 2522–2529.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36.

Jung, S., Qin, X., Oh, C., 2016. Improving strategic policies for pedestrian safety enhancement using classification tree modeling. Transp. Res. Part A: Policy Pract. 85, 53–64.

Kalra, N., Paddock, S.M., 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transp. Res. Part A: Policy Pract. 94, 182–193.

Ke, J., Zhang, S., Yang, H., Chen, X., 2019. PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. Transportmetrica A: Transp. Sci. 15 (2), 872–895.

Kleinbaum, D.G., Klein, M., 2010. Introduction to logistic regression. In: Logistic Regression. Springer, pp. 1–39.
Kondermann, D., Nair, R., Meister, S., Mischler, W., Güssefeld, B., Honauer, K., Hofmann, S., Brenner, C., Jähne, B., 2014. Stereo ground truth with error bars. In: Asian Conference on Computer Vision. Springer, pp. 595–610.
Koopman, P., Wagner, M., 2018. Toward a Framework for Highly Automated Vehicle Safety Validation. Technical Report, SAE Technical Paper.
Kuang, L., Yan, H., Zhu, Y., Tu, S., Fan, X., 2019. Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor. J. Intelligent Transportation Systems 23 (2), 161–174.
Kuhn, M., et al., 2008. Building predictive models in r using the caret package. J Stat Softw 28 (5), 1–26.
Kuhn, M., et al., 2013. Predictive modeling with R and the caret package. Google Scholar.
Kumar, K., Parida, M., Katiyar, V.K., 2015. Short term traffic flow prediction in heterogeneous condition using artificial neural network. Transport 30 (4), 397–405.
Leddar, 2021. Leddar PixSet dataset. https://leddartech.com/solutions/leddar-pixset-dataset/, Online; accessed 14 Dec 2021.
Leilabadi, S.H., Schmidt, S., 2019. In-depth analysis of autonomous vehicle collisions in california. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, pp. 889–893.
Level 5, 2021. Level 5 open data. https://level-5.global/data/, Online; accessed 14 Dec 2021.
Li, R., Zhai, R., 2019. Estimation and analysis of minimum traveling distance in self-driving vehicle to prove their safety on road test. J. Phys.: Conf. Ser. 1168 (3), 032101.
Liang, G., Zhang, C., 2012. An efficient and simple under-sampling technique for imbalanced time series classification. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 2339–2342.
Ligocki, A., Jelinek, A., Zalud, L., 2020. Brno urban dataset-the new data for self-driving agents and mapping tasks. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 3284–3290.
Liu, F., Zhao, F., Liu, Z., Hao, H., 2019. Can autonomous vehicle reduce greenhouse gas emissions? A country-level evaluation. Energy Policy 132, 462–473.
Liu, X.Y., Zhou, Z.H., 2006. The influence of class imbalance on cost-sensitive learning: An empirical study. In: Proceedings of the Sixth International Conference on Data Mining. IEEE, pp. 970–974.
Maddern, W., Pascoe, G., Linegar, C., Newman, P., 2017. 1 year, 1000 km: The oxford robotcar dataset. Int. J. Robot. Res. 36 (1), 3–15.
Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al., 2021. One million scenes for autonomous driving: Once dataset. arXiv preprint arXiv:2106.11037.
Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, S., Voit, M., Stiefelhagen, R., 2019. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2801–2810.
Mathias, M., Timofte, R., Benenson, R., Van Gool, L., 2013. Traffic sign recognition – How far are we from the solution? In: The 2013 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.
Meister, S., Jähne, B., Kondermann, D., 2012. Outdoor stereo camera system for the generation of real-world benchmark data sets. Opt. Eng. 51 (02), 021107.
Mujalli, R.O., de Ona, J., 2013. Injury severity models for motor vehicle accidents: a review. In: Proceedings of the Institution of Civil Engineers-Transport, Vol. 166. (5), Thomas Telford Ltd, pp. 255–270.
Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P., 2017. The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4990–4999.
Noh, S., 2018. Decision-making framework for autonomous driving at road intersections: Safeguarding against collision, overly conservative behavior, and violation vehicles. IEEE Trans. Ind. Electron. 66 (4), 3275–3286.
NSTC, U., 2020. Ensuring American leadership in automated vehicle technologies: Automated vehicles 4.0. Las Vegas. Recuperado El 25, 2020-2002.
NTSB, 2021. Investigation reports. https://www.ntsb.gov/investigations/AccidentReports/Pages/Reports.aspx?mode=Highway, Online; accessed 12 Dec 2021.
Palazzi, A., Abati, D., Calderara, S., Solera, F., Cucchiara, R., 2018. Predicting the driver's focus of attention: the DR(eye)VE project. IEEE Trans. Pattern Anal. Mach. Intell..
Pan, X., Shi, J., Luo, P., Wang, X., Tang, X., 2018. Spatial as deep: Spatial cnn for traffic scene understanding. In: Thirty-Second AAAI Conference on Artificial Intelligence.
Pandey, G., McBride, J.R., Eustice, R.M., 2011. Ford campus vision and lidar data set. Int. J. Robot. Res. 30 (13), 1543–1552.
Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A.K., 2019. Real-time accident detection: Coping with imbalanced data. Accid. Anal. Prev. 129, 202–210.
Pham, T.A., Yoo, M., 2020. Nighttime vehicle detection and tracking with occlusion handling by pairing headlights and taillights. Appl. Sci. 10 (11), 3986.
Prati, G., Pietrantoni, L., Fraboni, F., 2017. Using data mining techniques to predict the severity of bicycle crashes. Accid. Anal. Prev. 101, 44–54.
Pugeault, N., Bowden, R., 2015. How much of driving is preattentive? IEEE Trans. Veh. Technol. 64 (12), 5424–5438.
Rahimi, A., Azimi, G., Asgari, H., Jin, X., 2020. Injury severity of pedestrian and bicyclist crashes involving large trucks. In: International Conference on Transportation and Development 2020. American Society of Civil Engineers Reston, VA, pp. 110–122.
Rasouli, A., Kotseruba, I., Tsotsos, J.K., 2017. Agreeing to cross: How drivers and pedestrians communicate. In: IEEE Intelligent Vehicles Symposium (IV). pp. 264–269.
Reinhard Klette, 2021. EISATS. https://ccv.wordpress.fos.auckland.ac.nz/eisats/, Online; accessed 14 Dec 2021.
Ripley, B.D., 2007. Pattern Recognition and Neural Networks. Cambridge University Press.
Romera, E., Bergasa, L.M., Arroyo, R., 2016. Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 387–392.
Salon, D., McIntyre, A., 2018. Determinants of pedestrian and bicyclist crash severity by party at fault in San Francisco, CA. Accid. Anal. Prev. 110, 149–160.
Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Netw. 61, 85–117.
Schoettle, B., Sivak, M., 2015. A preliminary analysis of real-world crashes involving self-driving vehicles. Univ. Michigan Transp. Res. Inst..
Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A., 2008. A comparative study of data sampling and cost sensitive learning. In: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops. IEEE, pp. 46–52.
Sezer, V., Bandyopadhyay, T., Rus, D., Frazzoli, E., Hsu, D., 2015. Towards autonomous navigation of unsignalized intersections under uncertainty of human driver intent. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 3578–3585.
Shi, X., Wong, Y.D., Li, M.Z.F., Palanisamy, C., Chai, C., 2019. A feature learning approach based on XGBoost for driving assessment and risk prediction. Accid. Anal. Prev. 129, 170–179.
Song, Y., Chitturi, M.V., Noyce, D.A., 2021. Automated vehicle crash sequences: Patterns and potential uses in safety testing. Accid. Anal. Prev. 153, 106017.
Tarmizi, I.A., Abd Aziz, A., 2018. Vehicle detection using convolutional neural network for autonomous vehicles. In: 2018 International Conference on Intelligent and Advanced System (ICIAS). IEEE, pp. 1–5.
Teichman, A., Levinson, J., Thrun, S., 2011. Towards 3D object recognition via classification of arbitrary object tracks. In: 2011 IEEE International Conference on Robotics and Automation. IEEE, pp. 4034–4041.
Theofilatos, A., Antoniou, C., Yannis, G., 2021. Exploring injury severity of children and adolescents involved in traffic crashes in Greece. J. Traffic and Transportation Engineering 8 (4), 596–604.
Therneau, T.M., Atkinson, E.J., et al., 2019. An Introduction to Recursive Partitioning Using the RPART Routines. Technical Report, Technical report Mayo Foundation.
Ting, K.M., 1998. Inducing cost-sensitive trees via instance weighting. In: European Symposium on Principles of Data Mining and Knowledge Discovery. Springer, pp. 139–147.
Udacity, 2021. Udacity self-driving car dataset. https://github.com/udacity/self-driving-car, Online; accessed 13 Dec 2021.
Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C., 2019. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1743–1751.
Wang, S., Li, Z., 2019a. Exploring causes and effects of automated vehicle disengagement using statistical modeling and classification tree based on field test data. Accid. Anal. Prev. 129, 44–54.
Wang, S., Li, Z., 2019b. Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches. PLoS One 14 (3), e0214550.
Wang, S., Li, Z., 2019c. Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches. PLoS One 14 (3), e0214550.
Wang, H., Yu, Y., Cai, Y., Chen, X., Chen, L., Liu, Q., 2019. A comparative study of state-of-the-art deep learning algorithms for vehicle detection. IEEE Intell. Transp. Syst. Mag. 11 (2), 82–95.
Waymo, 2021a. Safety. https://waymo.com/safety/, Online; accessed 11 Dec 2021.
Waymo, 2021b. Waymo open dataset. https://waymo.com/open/#, Online; accessed 13 Dec 2021.
Weiss, G.M., McCarthy, K., Zabar, B., 2007. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? Dmin 7 (35–41), 24.
Weng, J., Meng, Q., 2012. Effects of environment, vehicle and driver characteristics on risky driving behavior at work zones. Saf. Sci. 50 (4), 1034–1042.
Wigness, M., Eum, S., Rogers, J.G., Han, D., Kwon, H., 2019. A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 5000–5007.
Wiseman, S., Adler-Golden, S., Ientilucci, E., Perkins, T., 2021. Enhanced target detection under poorly illuminated conditions. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, pp. 1425–1428.
Wojek, C., Walk, S., Schiele, B., 2009. Multi-cue onboard pedestrian detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 794–801.

Xu, C., Ding, Z., Wang, C., Li, Z., 2019. Statistical analysis of the patterns and characteristics of connected and autonomous vehicle involved crashes. J. Saf. Res. 71, 41–47.

Yahaya, M., Guo, R., Jiang, X., Bashir, K., Matara, C., Xu, S., 2021. Ensemble-based model selection for imbalanced data to investigate the contributing factors to multiple fatality road crashes in Ghana. Accid. Anal. Prev. 151, 105851.

Yan, X., Ma, M., Huang, H., Abdel-Aty, M., Wu, C., 2011. Motor vehicle–bicycle crashes in Beijing: Irregular maneuvers, crash patterns, and injury severity. Accid. Anal. Prev. 43 (5), 1751–1758.

Yan, X., Richards, S., Su, X., 2010. Using hierarchical tree-based regression model to predict train–vehicle crashes at passive highway-rail grade crossings. Accid. Anal. Prev. 42 (1), 64–74.

Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B., 2019. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 899–908.

Yogamani, S., Hughes, C., Horgan, J., Sistu, G., Varley, P., O'Dea, D., Uricár, M., Milz, S., Simon, M., Amende, K., et al., 2019. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9308–9318.

You, J., Wang, J., Guo, J., 2017. Real-time crash prediction on freeways using data mining and emerging techniques. J. Modern Transp. 25 (2), 116–123.

Yuan, Y., Sester, M., 2021. COMAP: A synthetic dataset for collective multi-agent perception of autonomous driving. Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci. 43, 255–263.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transp. Res. C 58, 308–324.

Zhu, S., Wan, J., 2021. Cost-sensitive learning for semi-supervised hit-and-run analysis. Accid. Anal. Prev. 158, 106199.

Zipfl, M., Fleck, T., Zofka, M.R., Zöllner, J.M., 2020. From traffic sensor data to semantic traffic descriptions: The test area autonomous driving Baden-Württemberg dataset (TAF-BW dataset). In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 1–7.